

Static Parameter Estimation for ABC Approximations of Hidden Markov Models

BY ELENA EHRLICH¹, AJAY JASRA² & NIKOLAS KANTAS³

¹Department of Mathematics, Imperial College London, London, SW7 2AZ, UK.

E-Mail: *elena.ehrlich05@ic.ac.uk*

²Department of Statistics & Applied Probability, National University of Singapore, Singapore, 117546, SG.

E-Mail: *staja@nus.edu.sg*

³Department of Statistical Science, University College London, London, WC1E 6BT, UK.

E-Mail: *n.kantas@ucl.ac.uk*

Abstract

In this article we focus on Maximum Likelihood estimation (MLE) for the static parameters of hidden Markov models (HMMs). We will consider the case where one cannot or does not want to compute the conditional likelihood density of the observation given the hidden state because of increased computational complexity or analytical intractability. Instead we will assume that one may obtain samples from this conditional likelihood and hence use approximate Bayesian computation (ABC) approximations of the original HMM. ABC approximations are biased, but the bias can be controlled to arbitrary precision via a parameter $\epsilon > 0$; the bias typically goes to zero as $\epsilon \searrow 0$. We first establish that the bias in the log-likelihood and gradient of the log-likelihood of the ABC approximation, for a fixed batch of data, is no worse than $\mathcal{O}(n\epsilon)$, n being the number of data; hence, for computational reasons, one might expect reasonable parameter estimates using such an ABC approximation. Turning to the computational problem of estimating θ , we propose, using the ABC-sequential Monte Carlo (SMC) algorithm in [18], an approach based upon simultaneous perturbation stochastic approximation (SPSA). Our method is investigated on two numerical examples.

Key-Words: Approximate Bayesian Computation, Hidden Markov Models, Parameter Estimation, Sequential Monte Carlo

1 Introduction

Hidden Markov models provide a flexible description of a wide variety of real-life phenomena; see [5] for an overview. An HMM is a pair of discrete-time stochastic processes, $\{X_n\}_{n \geq 0}$ and $\{Y_n\}_{n \geq 1}$, where $X_n \in \mathbf{X} \subseteq \mathbb{R}^{d_x}$ is an unobserved process and $y_n \in \mathbf{Y} \subseteq \mathbb{R}^{d_y}$ is observed. The hidden process $\{X_n\}_{n \geq 0}$ is a Markov chain with initial density $\mu_\theta(x_0)$ at time 0 and transition density $f_\theta(x_n|x_{n-1})$, with $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ i.e.

$$\mathbb{P}_\theta(X_0 \in A) = \int_A \mu_\theta(x_0) dx_0 \quad \text{and} \quad \mathbb{P}_\theta(X_n \in A | X_{n-1} = x_{n-1}) = \int_A f_\theta(x_n | x_{n-1}) dx_n \quad n \geq 1 \quad (1)$$

where \mathbb{P}_θ denotes probability, $A \in \mathcal{B}(\mathbf{X})$ and dx_n is Lebesgue measure. In addition, the observations $\{Y_n\}_{n \geq 1}$ conditioned upon $\{X_n\}_{n \geq 0}$ are statistically independent and have marginal density $g_\theta(y_n|x_n)$, i.e.

$$\mathbb{P}_\theta(Y_n \in B | \{X_k\}_{k \geq 0} = \{x_k\}_{k \geq 0}) = \int_B g_\theta(y_n | x_n) dy_n \quad n \geq 1 \quad (2)$$

with $B \in \mathcal{B}(\mathbf{Y})$. The HMM is given by equations (1)-(2) and is often referred to in the literature as a state-space model. Here θ is a static parameter, which is to be estimated in using MLE and online as the data arrive; this problem has a large range of real applications such as financial modelling or weather prediction.

Statistical inference from the class of HMMs described above is typically non-trivial. In most scenarios of practical interest one cannot calculate the likelihood:

$$p_\theta(y_{1:n}) = \int g_\theta(y_n | x_n) \pi_\theta(x_n | y_{1:n-1}) dx_n$$

where $y_{1:n} := (y_1, \dots, y_n)$ and $\pi_\theta(x_n | y_{1:n-1})$ is the predictor; see e.g. [5] for the standard filtering recursions. Hence as the likelihood is not analytically tractable, one must resort to numerical methods, not only to compute it, but to maximize $p_\theta(y_{1:n})$ w.r.t. θ . When θ is known, a popular collection of techniques for both estimating the likelihood as well as performing filtering and smoothing are sequential Monte Carlo methods e.g. [15]. SMC techniques simulate a collection of N samples in parallel, sequentially in time and combine importance sampling and resampling to approximate a sequence of probability distributions of increasing state-space known pointwise

up-to a multiplicative constant. These techniques provide a natural estimate of the likelihood. The estimate is quite well understood and is known to be unbiased [10] and, in addition, the relative variance is known to increase linearly with n [7, 30], n being the number of data. When θ is unknown, as is the case here, estimation of θ is complicated by the path-degeneracy problem of SMC methods; e.g. [26]. However, there are still many specialized SMC techniques which can successfully be used for online parameter estimation of HMMs in a wide variety of contexts, such as [13, 26]. Most of these techniques require the evaluation of $g_\theta(y|x)$ and potentially the gradient vectors as well.

In this article, we consider the scenario where $g_\theta(y_n|x_n)$ is either intractable, in the sense that one cannot calculate it or an unbiased estimator of it, or one does not want to calculate the density, potentially due to the high-dimensionality of X_n . It is assumed that one can sample from $g_\theta(y_n|x_n)dy_n$. In this case, one cannot use standard or the more advanced SMC methods that are mentioned above (or indeed many other techniques) and hence exact online parameter estimation is difficult to achieve. One approach which is designed to deal with this problem are ABC techniques; see e.g. [20]. Whilst there are a number of other competitors [17], we focus upon ABC ideas; see [17, 18] for some discussion of the relative merits of ABC against competing methods. In the context of HMMs there has been some work on the construction of ABC approximations of HMMs [18, 22], computational techniques for filtering and smoothing [18, 21, 6] and their statistical consistency for parameter estimation [8, 9]. ABC approximations of HMMs are biased, but the bias can be controlled to arbitrary precision via a parameter $\epsilon > 0$; the bias typically goes to zero as $\epsilon \searrow 0$. At present there is not a methodology which can achieve our objective of online parameter estimation. In this article we do the following:

1. Investigate the bias in the log-likelihood and the gradient of the log-likelihood that is induced by the ABC approximation for a fixed data set.
2. Develop an SMC approach with cost $\mathcal{O}(N)$ that allows one to estimate the static parameters in an online fashion.

In order to estimate the parameters one must obtain numerical estimates of the log-likelihood and gradient of this quantity. It is then important to understand what happens to the bias of the ABC approximation of these latter quantities, as the time parameter (number of data, n) grows. We establish, under some assumptions, that this ABC bias, for both quantities is no worse than $\mathcal{O}(n\epsilon)$; this result is associated to the theoretical work in [8, 9]. These former results indicate that the ABC approximation is amenable to numerical implementation: parameter estimation will not necessarily be dominated by the bias; we discuss why this is the case in Remarks 2.1 and 2.2. For 2. we introduce an SMC approach based upon SPSA [28] to estimate the parameters in an online manner (see also [27] in the context of HMMs). This methodology can be expected to ‘work well’ when:

- d_x is large and d_θ, d_y are small to moderate.

Whilst these statements are somewhat delicate (e.g. what is large), in the scenario of high-dimensional states, it has been established in [3] that the *simulation* error does not explode in the dimension. As a result, the ideas here can be seen as principled competitors (and related to - see [24]) to ensemble kalman filter-based algorithms such as in [16].

This paper is structured as follows. In Section 2 we discuss the model and ABC approximation. Our bias result is also given. In Section 3 our computational strategy is outlined. In Section 4 the method is investigated from a computational perspective. In Section 5 the article is concluded with some discussion of future work. The proofs of our results can be found in the appendix.

2 Model and Approximation

2.1 Model and Estimation

Consider first the joint filtering or smoothing density of the HMM given by

$$\pi_\theta(x_{0:n}|y_{1:n}) = \frac{\mu_\theta(x_0) \prod_{k=1}^n g_\theta(y_k|x_k) f_\theta(x_k|x_{k-1})}{\int_{X^{n+1}} \mu_\theta(x_0) \prod_{k=1}^n g_\theta(y_k|x_k) f_\theta(x_k|x_{k-1}) dx_{0:n}}$$

where $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ is the static parameter, $x_n \in \mathbf{X}$ are the hidden states and $y_n \in \mathbf{Y}$ the observations. This quantity can be computed recursively using

$$\pi_\theta(x_{0:k}|y_{1:k-1}) = \int_{\mathbf{X}} \pi_\theta(x_{0:k-1}|y_{1:k-1}) f_\theta(x_k|x_{k-1}) dx_k \quad (3)$$

$$\pi_\theta(x_{0:k}|y_{1:k}) = \frac{g_\theta(y_k|x_k) \pi_\theta(x_{0:k}|y_{1:k-1})}{p_\theta(y_k|y_{1:k-1})} \quad (4)$$

with the *recursive likelihood* being

$$p_\theta(y_k|y_{1:k-1}) = \int_{\mathbf{X}} g_\theta(y_k|x_k) \pi_\theta(x_{0:k}|y_{1:k-1}) dx_k \quad (5)$$

Furthermore we write the log- (marginal) likelihood at time n :

$$\log(p_\theta(y_{1:n})) = \sum_{k=1}^n \log(p_\theta(y_k|y_{1:k-1})).$$

In the context of MLE one is usually interested computing

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \log(p_\theta(y_{1:n}))$$

Note that this is a batch or off-line method, which means that one needs to wait first to collect the complete dataset and then compute the ML estimate. For a long observation sequence the computation of the gradient at each iteration of the algorithm can be prohibitive. Therefore, one uses on-line methods whereby the estimate of the parameter is updated sequentially as the data arrives. A practical alternative would be to consider the following update scheme at time k , for some sequence $\{a_k\}_{k \geq 1}$

$$\theta_{k+1} = \theta_k + a_{k+1} \nabla \log(p_\theta(y_k|y_{1:k-1}))|_{\theta=\theta_k}.$$

Upon receiving y_k , the parameter estimate is updated in the direction of ascent of the conditional density of this new observation. The algorithm in the present form is not suitable for on-line implementation due to the need to evaluate the gradient of $\log p_\theta(y_k|y_{0:k-1})$ at the current parameter estimate which would require computing the filter from time 0 to time k using the current parameter value θ_k .

A recursive ML (RML) algorithm bypassing this problem has been proposed in the literature when \mathbf{X} is finite in [19]. It relies on the following update scheme

$$\theta_{k+1} = \theta_k + a_{k+1} \nabla \log(p_{\theta_{0:k}}(y_k|y_{1:k-1}))$$

where the positive non-increasing step-size sequence $\{a_k\}_{k \geq 1}$ satisfies $\sum_k a_k = \infty$ and $\sum_k a_k^2 < \infty$ [19]; e.g. $a_k = k^{-\alpha}$ for $0.5 < \alpha \leq 1$. The quantity $\nabla \log p_{\theta_{0:k}}(y_k|y_{1:k-1})$ is defined as

$$\nabla \log(p_{\theta_{0:k}}(y_k|y_{1:k-1})) = \nabla \log(p_{\theta_{0:k}}(y_{1:k})) - \nabla \log(p_{\theta_{0:k-1}}(y_{1:k-1}))$$

where the notation $\nabla \log(p_{\theta_{0:k}}(y_{1:k}))$ indicates that at each time k the quantities in (3)-(5) are computed using the parameter estimate θ_k . The asymptotic properties of this algorithm (i.e. the behavior of θ_k in the limit as k goes to infinity) have been studied in [19] for a finite state-space HMM. It is shown that under regularity conditions this algorithm converges towards a local maximum of the average log-likelihood; this average log-likelihood being maximized at the ‘true’ parameter value.

In this article, we would like to implement approximate versions of these on-line and off-line ML schemes when both the following cases hold:

- Case 1: We can sample from the conditional distribution of $Y|x$, for any fixed θ and x .
- Case 2: We cannot or do not want to evaluate the conditional density of $Y|x$, $g_\theta(y|x)$ and do not have access to an unbiased estimate of it.

Apart from using likelihoods which do not admit computable densities such as some stable distributions, this context might appear relevant to the context when one is interested to use SMC methods and evaluate $g_\theta(y|x)$ when d_x is large. SMC methods for *filtering* do not always scale well with the dimension of the hidden state d_x , often requiring a computational cost $\mathcal{O}(\kappa^{d_x})$, with $\kappa > 1$; see e.g. [3, 4]. A more detailed discussion on the difficulties of using SMC methods in high dimensions is far beyond the scope of this article, but we remark the ideas in this paper can be relevant in this context.

2.2 ABC Approximation and Noisy ABC

To facilitate statistical inference, we consider an ABC approximation of the joint smoothing density (e.g. [18, 22]):

$$\pi_{\theta,\epsilon}(u_{1:n}, x_{0:n}|y_{1:n}) = \frac{\mu_{\theta}(x_0) \prod_{k=1}^n K_{\theta,\epsilon}(y_k|u_k) g_{\theta}(u_k|x_k) f_{\theta}(x_k|x_{k-1})}{\int_{\mathcal{X}^{n+1} \times \mathcal{Y}^n} \mu_{\theta}(x_0) \prod_{k=1}^n K_{\theta,\epsilon}(y_k|u_k) g_{\theta}(u_k|x_k) f_{\theta}(x_k|x_{k-1}) du_{1:n} x_{0:n}}$$

where $u_n \in \mathcal{Y}$ are pseudo observations, $\epsilon > 0$ $K_{\theta} : \mathcal{Y} \times \mathbb{R}_+ \times \Theta \rightarrow \mathbb{R}_+ \cup \{0\}$ is some kernel function that has bandwidth that depends upon a precision parameter $\epsilon > 0$. Examples include:

$$\begin{aligned} K_{\theta,\epsilon}(y_k|u_k) &= \mathbb{I}_{\{u:|y_k-u|<\epsilon\}}(u_k) \\ K_{\theta,\epsilon}(y_k|u_k) &= \phi_{d_y}(y_k; u_k, \epsilon I_{d_x}) \end{aligned}$$

where \mathbb{I} is the indicator function, $|\cdot|$ is the \mathbb{L}_1 -norm, $\phi_d(y; \xi, \Sigma)$ is normal density on d -dimensions with mean ξ and covariance Σ and I_d is the d -dimensional identity matrix.

Consider the quantity, to be used below:

$$g_{\theta,\epsilon}(y_k|x_k) = \frac{\int_{\mathcal{Y}} K_{\theta,\epsilon}(y_k|u_k) g_{\theta}(u_k|x_k) du_k}{\int_{\mathcal{Y}^2} K_{\theta,\epsilon}(y_k|u_k) g_{\theta}(u_k|x_k) du_k dy_k}. \quad (6)$$

Throughout the article we *critically* choose $K_{\theta,\epsilon}(y_k|u_k)$ such that the denominator of (6) does not depend upon x_k or θ . As noted in [8], after integrating out the $u_{1:n}$, this representation leads to a new (or perturbed) HMM with transitions f_{θ} and likelihoods $g_{\theta,\epsilon}$. Parameter estimation associated to the smoother $\pi_{\theta,\epsilon}$ just considers the function:

$$\log(p_{\theta,\epsilon}(y_{1:n})) = \sum_{k=1}^n \log(p_{\theta,\epsilon}(y_k|y_{1:k-1}))$$

where

$$p_{\theta,\epsilon}(y_k|y_{1:k-1}) = \int_{\mathcal{X}} g_{\theta,\epsilon}(y_k|x_k) \pi_{\theta,\epsilon}(x_k|y_{1:k-1}) dx_k.$$

We term the maximizer of $p_{\theta,\epsilon}(y_{1:n})$ as the ABC-MLE. One can then define a RML procedure for the ABC-HMM as in Section 2.1:

$$\theta_{k+1} = \theta_k + a_{k+1} \nabla \log\{p_{\theta,\epsilon}(y_k|y_{1:k-1})\}.$$

In practice, one can consider an estimation of $p_{\theta,\epsilon}(y_{1:n})$ including factors independent of θ, ϵ ; this is discussed in Section 3.

Results on associated to the asymptotics of the ABC-MLE (i.e. as n grows) can be found in [8, 9]; there is an asymptotic bias. In addition, in the case of noisy ABC, where the data become corrupted, there is no asymptotic bias and one can recover the true parameter. We remark that the methodology that is considered in this article can *easily* incorporate noisy ABC. However, there may be several reasons why one may not want to use noisy ABC: (1) the consistency results (currently) depend upon the data originating from the original HMM; (2) the current simulation-based methodology may not be able to push ϵ towards zero. For (1), if the data do not originate from the HMM of interest, it has not been studied what happens with regards to the asymptotics of noisy ABC for HMMs. It may be that some investigators might be uncomfortable with assuming that the data originate from the exactly the HMM being fitted. For (2) the asymptotic bias (which is under assumptions either $\mathcal{O}(\epsilon)$ or $\mathcal{O}(\epsilon^2)$ [8, 9]) could be less than the asymptotic variance (under assumptions $\mathcal{O}(\epsilon^2)$ [8, 9]) as ϵ could be much bigger than 1 when using current simulation methodology. We do not use noisy ABC in this article, but acknowledge its fundamental importance with regards to parameter estimation associated to ABC for HMMs; our approach is pragmatic, taking into account points (1)-(2).

2.3 Result

We now prove an upper-bound on the bias induced by the ABC approximation on the log-likelihood and gradient of the log-likelihood. The latter is more relevant for parameter estimation, but the mathematical arguments are considerably more involved for this quantity, in comparison to the ABC bias of the log-likelihood. Hence the log-likelihood is considered as a simple preliminary result. These results are to be taken in the context of ABC (not noisy ABC) and help to provide some guarantees associated to the numerics.

We consider the scenario

$$K_{\theta,\epsilon}(y_k|u_k) = \mathbb{I}_{A_{\epsilon,y_k}}(u_k)$$

where the set A_{ϵ, y_k} is specified below. Throughout $|\cdot|$ is understood to be an \mathbb{L}_1 -norm. The hidden-state is assumed to lie on a *compact* set, i.e. \mathbf{X} is compact. We use the notation $\mathcal{P}(\mathbf{X})$ to denote the class of probability measures on \mathbf{X} and $\mathcal{M}(\mathbf{X})$ the collection of finite and signed measures on \mathbf{X} . $\|\cdot\|$ denotes the total variation distance. The initial distribution of the hidden Markov chain is written as $\mu_\theta \in \mathcal{P}(\mathbf{X})$. In addition, we condition on the observed data and do not mention them in any mathematical statement of results (due to the assumptions below). We do not consider the instance of whether the data originate, or not, from a HMM. For the control of the bias of the gradient of the log-likelihood (Theorem 2.1), we assume that $d_\theta = 1$. This is not restrictive as one can use the arguments to prove analogous results when $d_\theta > 1$, by considering componentwise arguments for the gradient. In addition, for the gradient result, the derivative of μ_θ is written $\widetilde{\mu}_\theta \in \mathcal{M}(\mathbf{X})$. We make the following assumptions, which are extremely strong. They are made to keep the proofs as short as possible.

(A1) *Lipschitz Continuity of the Likelihood.* There exist $L < +\infty$ such that for any $x \in \mathbf{X}$, $y, y' \in \mathbf{Y}$, $\theta \in \Theta$

$$|g_\theta(y|x) - g_\theta(y'|x)| \leq L|y - y'|.$$

(A2) *Statistic and Metric.* The set $A_{\epsilon, y}$ is:

$$A_{\epsilon, y} = \{u : |y - u| < \epsilon\}.$$

(A3) *Boundedness of Likelihood and Transition.* There exist $0 < \underline{C} < \overline{C} < +\infty$ such that for all $x, x' \in \mathbf{X}$, $y \in \mathbf{Y}$, $\theta \in \Theta$

$$\begin{aligned} \underline{C} &\leq f_\theta(x'|x) \leq \overline{C}, \\ \underline{C} &\leq g_\theta(y|x) \leq \overline{C}. \end{aligned}$$

(A4) *Lipschitz Continuity of the Gradient of the Likelihood.* $f_\theta(x'|x)$, $g_\theta(y|x')$ are differentiable in θ for each $x, x' \in \mathbf{X}$, $y \in \mathbf{Y}$. In addition, there exist $L < +\infty$ such that for any $x \in \mathbf{X}$, $y, y' \in \mathbf{Y}$, $\theta \in \Theta$

$$|\nabla\{g_\theta(y|x)\} - \nabla\{g_\theta(y'|x)\}| \leq L|y - y'|.$$

(A5) *Boundedness of Gradients of the Likelihood and Transition.* There exist $0 < \underline{C} < \overline{C} < +\infty$ such that for all $x, x' \in \mathbf{X}$, $y \in \mathbf{Y}$, $\theta \in \Theta$

$$\begin{aligned} \underline{C} &\leq \nabla\{f_\theta(x'|x)\} \leq \overline{C}, \\ \underline{C} &\leq \nabla\{g_\theta(y|x)\} \leq \overline{C}. \end{aligned}$$

We first have the result on the ABC bias of the log-likelihood. The proof is in appendix B.

Proposition 2.1. *Assume (A1-3). Then there exist a $C < +\infty$ such that for any $n \geq 1$, $\mu_\theta \in \mathcal{P}(\mathbf{X})$, $\epsilon > 0$, $\theta \in \Theta$ we have:*

$$|\log(p_\theta(y_{1:n})) - \log(p_{\theta, \epsilon}(y_{1:n}))| \leq Cn\epsilon.$$

Remark 2.1. *The above proposition gives some simple guarantees on the bias of the ABC log-likelihood. When using SMC algorithms to approximate $\log(p_\theta(y_{1:n}))$, the overall error will be decomposed into the deterministic bias that is present from the ABC approximation (that in Proposition 2.1) and the numerical error of approximating the log-likelihood. Under some assumptions, the \mathbb{L}_2 -error of the SMC estimate of the log-likelihood should not deteriorate any faster than linearly in time; this is due to the results cited previously. Thus, as the time parameter increases, the ABC bias of the log-likelihood will not necessarily dominate the simulation-based error that would be present even if g_θ is evaluated.*

Proposition 2.1 is reasonably straight-forward to prove, but, is of less interest in the context of parameter estimation, as one is interested in the gradient of the log-likelihood. We now have the result on the ABC bias of the gradient of the log-likelihood. The proof in appendix C.

Theorem 2.1. *Assume (A1-5). Then there exist a $C < +\infty$ such that for any $n \geq 1$, $\mu_\theta \in \mathcal{P}(\mathbf{X})$, $\widetilde{\mu}_\theta \in \mathcal{M}(\mathbf{X})$, $\epsilon > 0$, $\theta \in \Theta$ we have:*

$$|\nabla\{\log(p_\theta(y_{1:n}))\} - \nabla\{\log(p_{\theta, \epsilon}(y_{1:n}))\}| \leq Cn\epsilon(2 + \|\widetilde{\mu}_\theta\|).$$

Remark 2.2. The above Theorem again provides some explicit guarantees when using an ABC approximation along with SMC-based numerical methods. For example, if one can consider approximating gradients in an ABC context (see [31]), then from the results of [14], one expects that the variance of the SMC estimates to increase only linearly in time. Again, as time increases the ABC bias does not necessarily dominate the variance that would be present even if g_θ is evaluated (i.e. one uses SMC on the true model).

Remark 2.3. The result in Theorem 2.1 can be found in eq. (72) of [8] and direct limit (as $\epsilon \searrow 0$) in [9]. However, we adopt a new (and fundamentally different) proof technique, with a substantially clearer proof and an additional result of independent interest is proved. We derive the stability w.r.t. time of the bias of the ABC approximation of the filter derivative; see Theorem D.1 in appendix D.

3 Computational Strategy

3.1 SMC

In order to perform online parameter estimation, we will need to use a SMC algorithm to approximate $p_{\theta,\epsilon}(y_k|y_{1:k-1})$ for θ fixed; this is a critical quantity that we will use below. An algorithm which can do this is the SMC approach in [18] which is detailed in Figure 1, with proposals $\{q_{k,\theta}\}_{1 \leq k \leq n}$ with density w.r.t. Lebesgue measure.

On the basis of Figure 1, one can approximate $p_{\theta,\epsilon}(y_{1:n})$, *up-to a constant that is independent of θ* , as follows. In an abuse of notation, we denote this SMC estimate (which does not include factors that do not depend on θ) as $p_{\theta,\epsilon}^N(y_{1:n})$. The SMC estimate is

$$p_{\theta,\epsilon}^N(y_{1:n}) = \prod_{k=1}^n \frac{1}{N} \sum_{i=1}^N \widetilde{W}_k^{(i)}$$

with

$$p_{\theta,\epsilon}^N(y_k|y_{1:k-1}) = \frac{1}{N} \sum_{i=1}^N \widetilde{W}_k^{(i)}.$$

These estimates are unbiased for any $N \geq 1$ (see [10]). In practice, we are interested in the log-likelihoods; taking logarithms of the above estimates generally leads to a biased approximation of $\log\{p_{\theta,\epsilon}(y_{1:n})\}$ and $\log\{p_{\theta,\epsilon}(y_k|y_{1:k-1})\}$. One can implement a form of bias correction, using the Taylor series expansion ideas in [25]. Throughout, we use the bias-corrected estimates:

$$\begin{aligned} \log\{\widehat{p_{\theta,\epsilon}(y_{1:n})}\} &= \log\{p_{\theta,\epsilon}^N(y_{1:n})\} + \frac{1}{2N} p_{\theta,\epsilon}^N(y_{1:n})^{-2} \\ \log\{\widehat{p_{\theta,\epsilon}(y_k|y_{1:k-1})}\} &= \log\{p_{\theta,\epsilon}^N(y_k|y_{1:k-1})\} + \frac{1}{2N} p_{\theta,\epsilon}^N(y_k|y_{1:k-1})^{-2}. \end{aligned} \quad (7)$$

The parameter ϵ can be computed adaptively; see [18]. It is remarked that a drawback of this algorithm is that when d_y grows with ϵ, N fixed, one cannot expect the algorithm to work well for every ϵ ; typically one must increase ϵ to yield reasonable algorithmic results and this is at the cost of increasing the bias. To maintain ϵ at a reasonable level, one must consider more advanced strategies which are not investigated here.

One final point, which is often useful in practice. One can modify the ABC approximation to:

$$\pi_{\theta,\epsilon}(u_{1:n}^1, \dots, u_{1:n}^M, x_{0:n}|y_{1:n}) \propto \mu_\theta(x_0) \prod_{k=1}^n \left[\left(\frac{1}{M} \sum_{j=1}^M K_{\theta,\epsilon}(y_k|u_k^j) \right) \prod_{j=1}^M g_\theta(u_k^j|x_k) \right] f_\theta(x_k|x_{k-1})$$

which yields the same bias as the original ABC approximation (on integrating the u variables) but can yield substantial computational improvements. This is because as M grows one approximates a marginal SMC that does not sample the auxiliary u variables.

Remark 3.1. We note that, suppressing θ , if the HMM can be written in the form:

$$\begin{aligned} Y_n &= \xi_n(X_n, W_n) \quad n \geq 1 \\ X_n &= \varphi_n(X_{n-1}, V_n) \quad n \geq 1 \end{aligned}$$

where $X_0 = x_0 \in \mathsf{X}$ is known, $Y_n \in \mathsf{Y}$, $V_n \in \mathsf{X}$ with $\{V_n\}_{n \geq 1}$ i.i.d. $W_n \in \mathsf{Y}$ with $\{W_n\}_{n \geq 1}$ i.i.d. and independent of $\{V_n\}_{n \geq 1}$ and $\xi_n : \mathsf{X} \times \mathsf{Y} \rightarrow \mathsf{Y}$, $\varphi_n : \mathsf{X} \times \mathsf{X} \rightarrow \mathsf{X}$. Suppose that:

- Step 0. For $i = 1, \dots, N$ sample $X_0^{(i)}$ i.i.d. from $\mu_\theta(x_0)dx_0$. Set $W_0^{(i)} = 1/N$ for each $i \in \{1, \dots, N\}$. Set $k = 0$.
- Step 1. Resample N particles from

$$\hat{\pi}_k(\cdot) = \sum_{i=1}^N W_k^{(i)} \delta_{x_k^{(i)}}(\cdot),$$

which are also denoted $\{x_k^{(i)}\}$, and set $W_k^{(i)} = \frac{1}{N}$. Set $k = k + 1$ and if $k = n + 1$, stop.

- Step 2. For $i = 1, \dots, N$, sample $X_k^{(i)}$ from $q_{k,\theta}(x_k|x_{k-1}^{(i)})dx_k$ and $U_k^{(i)}$ from the likelihood $g_\theta(u_k|x_k^{(i)})du_k$. Compute

$$W_k^{(i)} \propto W_{k-1}^{(i)} \widetilde{W}_k^{(i)} \quad \widetilde{W}_k^{(i)} = \frac{K_{\theta,\epsilon}(y_k|u_k^{(i)})f_\theta(x_k^{(i)}|x_{k-1}^{(i)})}{q_{k,\theta}(x_k|x_{k-1}^{(i)})},$$

renormalize the weights and return to Step 1.

Figure 1: SMC Algorithm for ABC target.

- One can evaluate the densities of W_n and V_n and sample from the associated distributions.
- One can evaluate ξ_n (resp. φ_n) pointwise, for each $n \geq 1$ and X_n, W_n (resp. X_{n-1}, V_n).

One can construct a ‘collapsed’ (see [23]) ABC approximation (assuming $K_{\theta,\epsilon}(y|u) = \mathbb{I}_{A_{y,\epsilon}}(u)$, $A_{y,\epsilon} = \{u \in \mathcal{Y} : d(u, y) < \epsilon\}$, with d a distance metric on \mathcal{Y})

$$\pi_\epsilon(w_{1:n}, v_{1:n}|y_{1:n}) \propto \prod_{k=1}^n \mathbb{I}_{A_{y_k,\epsilon}}(\xi_k(\varphi^{(k)}(x_0, v_{1:k}), w_k))p(w_k)p(v_k).$$

Hence a version of the SMC algorithm in Figure 1 can be derived which does not need to sample from the dynamics of the data. In addition one does not need access to the transition density of the hidden Markov chain. This representation, however, does not always apply.

3.2 SPSA

Recall the RML procedure in Section 2.1, where $g_\theta(y|x)$ is not intractable:

$$\theta_{k+1} = \theta_k + a_{k+1} \nabla \log(p_{\theta_{0:k}}(y_k|y_{1:k-1})) \quad (8)$$

for $\{a_n\}$ a sequence of step-sizes. In practice, one does not know the gradient and must resort to (e.g.) SMC techniques to approximate it; see for example [26]. In our ABC context one can run the algorithm in Figure 1 to approximate the ABC filter. To recursively update θ , at least using the ideas in [26], one has to evaluate

$$\log(g_\theta(y|x)) \quad \text{and} \quad \nabla \log(g_\theta(y|x)) \quad (9)$$

which we will not have access to.

We propose the following computational scheme; the idea is to use SPSA, which does not require the quantities in (9). Introduce a decreasing sequence of positive numbers $\{c_k\}$. Suppose, with $\{a_k\}$ as in the update, (8), we have

$$\forall k, a_k > 0 \quad a_k, c_k \rightarrow 0 \quad \sum_k a_k = \infty \quad \sum_k \frac{a_k^2}{c_k^2} < \infty.$$

Start with some initial guess θ_0 and perform the standard SMC update (i.e. as in Figure 1) for two sets of particles. One with parameter:

$$\theta_0 + c_0 \Delta_0$$

and the other with parameter:

$$\theta_0 - c_0 \Delta_0$$

where Δ_0 is a d_θ -dimensional vector with each entry ± 1 Bernoulli distributed (see [28]). For both algorithms compute $\log(\widehat{p_{\theta_0+c_0\Delta_0,\epsilon}}(y_1))$ and $\log(\widehat{p_{\theta_0-c_0\Delta_0,\epsilon}}(y_1))$ respectively, where the estimates are the bias-corrected versions as in equation (7). To obtain the next parameter estimate, in the i^{th} -dimension, take

$$\theta_{1,i} = \theta_{0,i} + a_1 \frac{\log(\widehat{p_{\theta_0+c_0\Delta_0,\epsilon}}(y_1)) - \log(\widehat{p_{\theta_0-c_0\Delta_0,\epsilon}}(y_1))}{2c_0\Delta_{0,i}}.$$

At any subsequent time-point, with θ_k and perform the standard SMC update for two sets of particles. One with parameter:

$$\theta_k + c_k \Delta_k$$

and the other with parameter:

$$\theta_k - c_k \Delta_k$$

For both algorithms compute $\log(\widehat{p_{\theta_k+c_k\Delta_k,\epsilon}}(y_k|y_{1:k-1}))$ and $\log(\widehat{p_{\theta_k-c_k\Delta_k,\epsilon}}(y_k|y_{1:k-1}))$ To obtain the next parameter estimate, in the i^{th} -dimension, take

$$\theta_{k+1,i} = \theta_{k,i} + a_{k+1} \frac{\log(\widehat{p_{\theta_k+c_k\Delta_k,\epsilon}}(y_k|y_{1:k-1})) - \log(\widehat{p_{\theta_k-c_k\Delta_k,\epsilon}}(y_k|y_{1:k-1}))}{2c_k\Delta_{k,i}}.$$

This algorithm does not require one to evaluate g_θ or its gradient. We refer the reader to [28] and [27] for a theoretical justification of this procedure.

4 Numerical Simulations

We consider two numerical examples that are designed to investigate the accuracy and behaviour of our numerical algorithms. In order to do this, we do not consider scenarios where g_θ is intractable.

4.1 Linear Gaussian Model

We consider the following linear Gaussian HMM, with $\mathbf{Y} = \mathbf{X} = \mathbb{R}$:

$$\begin{aligned} Y_n &= X_n + \sigma_w W_n \\ X_n &= \phi X_{n-1} + \sigma_v V_n, \end{aligned}$$

with W_n, V_n independent and $W_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, $V_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. In the subsequent examples, we will use a simulated dataset obtained with $\theta = (\sigma_v, \phi, \sigma_w) = (0.2, 0.9, 0.3)$.

4.1.1 Offline MLE

We begin by considering a small data set, of $n = 1000$ data points. The offline scenario is the one for which we can expect the best possible performance of the ABC-SMC; if we cannot obtain reasonable parameter estimates in this scenario we would not expect ABC to be useful in practice. We are concerned with obtaining offline ABC-SMC estimates

$$\theta_{j+1}(i) = \theta_j(i) + a_{j+1} \frac{\log(\widehat{p_{\theta_j+c_j\Delta_j}}(y_{1:1000})) - \log(\widehat{p_{\theta_j-c_j\Delta_j}}(y_{1:n}))}{2c_j\Delta_j(i)},$$

where j is the iteration, $\theta_j(i)$ is the parameter estimate in the i^{th} -dimension, and $\Delta_j(i)$ is the i^{th} -entry of the Bernoulli distributed vector. For the SPSA stepsizes, we chose $c_j = j^{-0.1}$, $a_j = 1$ for $j < 10000$, and $a_j = (j - 10000)^{-0.8}$ for $j \geq 10000$. The iteration consists of running the ABC-SMC algorithm for 1000 data-points, with the current value of θ .

In Figure 2, we compare offline estimates of the following cases:

- (a) Kalman Filter (KF) with SPSA
- (b) SMC on the true model using $N = 1000$, with SPSA
- (c) ABC-SMC using $N = 200$, $M = 10$, $\epsilon = 0.1$, with SPSA
- (d) Maximum Likelihood estimates (MLE) from an offline grid search optimization.

In this particular test case, we can observe good relative performance of the ABC-SMC procedure, with regards to estimating parameters. This strong performance allows us to investigate a slightly more challenging scenario.

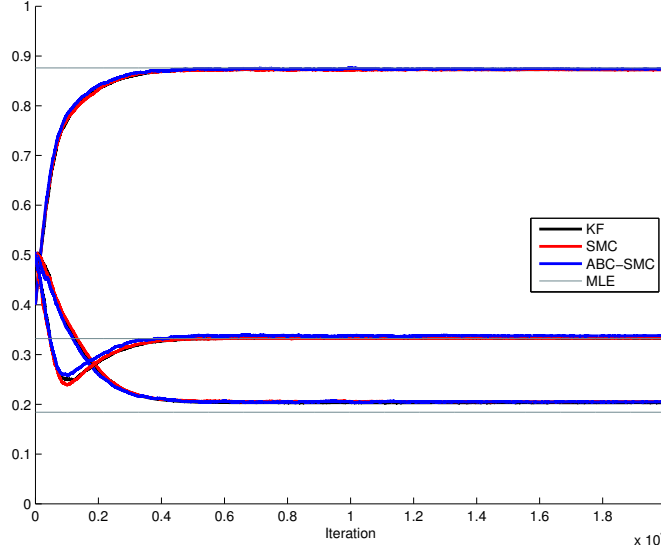


Figure 2: A typical run of the offline parameter estimates obtained by the KF, SMC, and ABC-SMC for the linear Gaussian HMM, along with their parameters' offline MLEs.

4.1.2 Online MLE

We now consider a larger data set with $n = 50,000$ data points, simulated with the previously indicated parameter values. We use the online SPSA method described in Section 3.2. The SMC (i.e. on the true model) and ABC-SMC algorithms were employed with the same N (and M , ϵ for ABC-SMC) as in the offline case, and the SPSA sequences are similar to their offline forms, in Section 4.1.1.

We ran fifty independent runs of the each algorithm considered in the previous Section. In Figure 3, we plot the medians and credible intervals for the 25-75% and 5-95% percentiles of the parameter estimates (across the independent runs). The $\hat{\theta}_k$ converge after $k = 20000$ time steps, with the KF and SMC yielding similarly valued estimates. We observe increased variance from left to right in Figure 3, which we attribute to the randomness of SMC and ABC-SMC respectively. In particular, the expected reduced accuracy of ABC-SMC against SMC is apparent, but, the bias does not appear to be substantial (for ABC-SMC) in this particular example.

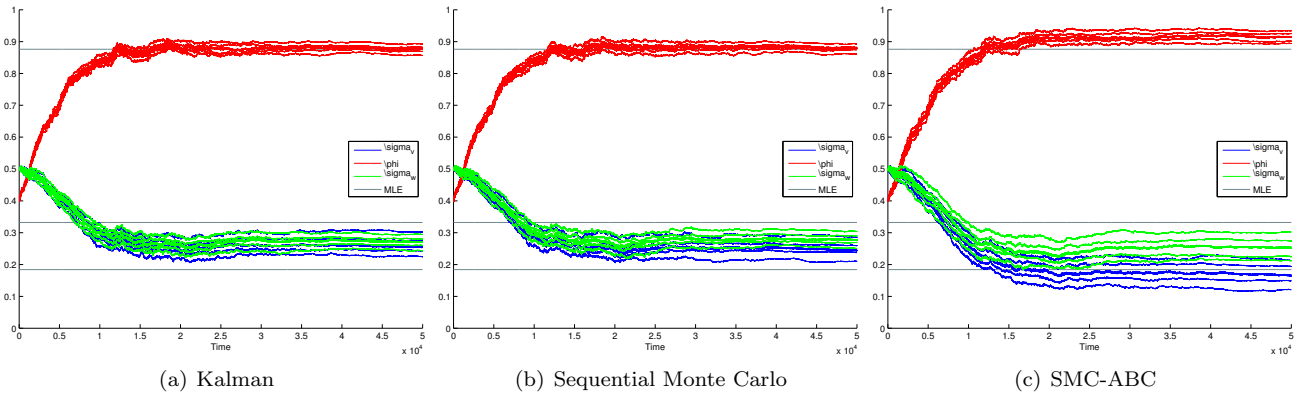


Figure 3: Credible intervals for the 5-95% and 25-75% percentiles, and the medians for multiple runs of online parameter estimates streamed by the KF, SMC, and ABC-SMC for the linear Gaussian HMM.

4.2 Lorenz '63 Model

4.2.1 Model and Data

We now consider the following non-linear state-space model with $\mathbf{X} = \mathbf{Y} = \mathbb{R}^3$. The original model is such that hidden process evolves deterministically according to the Lorenz '63 system of ordinary differential equations,

$$\begin{aligned}\frac{dX_t(1)}{dt} &= \sigma_{63}(X_t(2) - X_t(1)) \\ \frac{dX_t(2)}{dt} &= \rho X_t(1) - X_t(2) - X_t(1)X_t(3) \\ \frac{dX_t(3)}{dt} &= X_t(1)X_t(2) - \beta X_t(3).\end{aligned}$$

where we recall that the arguments $X_t(j)$ are the j^{th} -dimension at time t ; where t is continuous here. We modify the model to one such that the hidden process is a discrete-time Markov chain with stochastic dynamics:

$$X_n = f_n(X_{n-1}) + V_n, \quad n \geq 1$$

where f_n is the 4th-order approximation Runge Kutta solution to the Lorenz '63 system, $V_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau I_{d_x})$ and X_0 is taken as known. Here τ is used to represent the time-discretization.

For the observations:

$$Y_n = HX_n + QW_n, \quad n \geq 1$$

where $W_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_y})$, W_n is independent of V_n and Q is the Cholesky root of a Toeplitz matrix defined by the parameters κ and σ as follows:

$$Q_{ij} = \sigma S(\kappa^{-1} \min(|i-j|, d_y - |i-j|)), \quad i, j \in \{1, \dots, d_y\}$$

$$S(z) = \begin{cases} 1 - \frac{3}{2}z + \frac{1}{2}z^3, & 0 \leq z \leq 1 \\ 0, & z > 1 \end{cases},$$

and

$$H_{ij} = \begin{cases} \frac{1}{2}, & i = j \\ \frac{1}{2}, & i = j - 1 \\ 0, & i \neq j \end{cases}.$$

When $\theta = (\kappa, \sigma, \sigma_{63}, \rho, \beta) = (2.5, 2, 10, 28, \frac{8}{3})$, $n = 5000$ and $\tau = 0.05$, a visualisation of the Lorenz '63 (hidden) dynamics is shown in Figure 4(a) and the associated simulated dataset in 4(b).

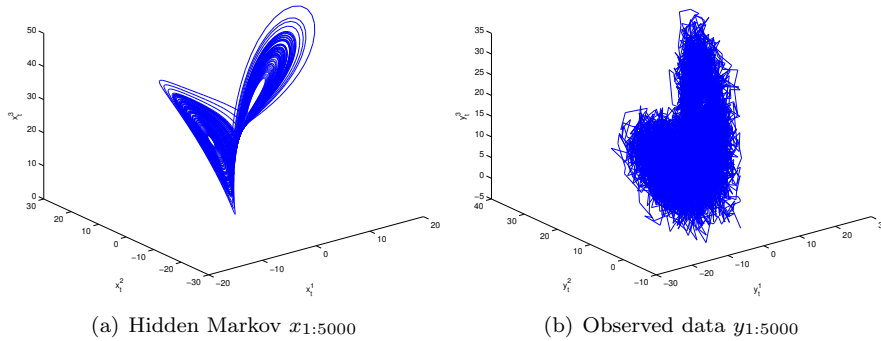


Figure 4: Evolution of the 3-dimensional Lorenz '63 HMM in Section 4.2

For the simulated dataset in Figure 4(b), we use ABC-SMC to obtain online parameter estimates for θ and we study the performance of these estimates under different settings. We will use $\hat{\theta}_{\epsilon, n}^{N, M}$ to denote the estimate of θ at time n , that was estimated using N particles, M pseudo-observations and a Gaussian kernel with covariance ϵI_{d_y} . We will compare the behaviour of the algorithm as each of N, M, n, ϵ varies.

4.2.2 Numerical Results

We now examine the performance of the algorithm with $N \in \{100, 1000, 10000\}$. For each value of N , we ran fifty independent runs of ABC-SMC, using $M = 10$ and $\epsilon = 1$. In Figures 5(a)-5(d) we plot boxplots of the terminal parameter estimates, $\hat{\theta}_{1,5000}^{N,10}$, against their true values marked by dotted green lines. In Figures 5(e)-5(h) we plot the absolute value of the Monte Carlo (MC) bias (that is, the absolute difference between the estimate and true value), in red, and the MC standard deviation, in blue. The MC bias and standard deviation points are fitted with least-squares curves proportional to $\frac{1}{\sqrt{N}}$, the standard MC rates with which the accuracy of the estimates is expected to improve. With regards to the variability of the estimates one sees the expected reduction in variability as N increases. The bias is harder to quantify; it will not necessarily be the case that as N grows the bias falls. This is because there is a Monte Carlo bias (from the SMC), an optimization bias (from the SPSA), an approximation bias (from the ABC) and the fact that the data have been generated from the model (so the true static parameters might not be exact). Increasing N can only deal with the SMC bias (which for estimates with parameters fixed is $\mathcal{O}(N^{-1})$), but the addition of parameter estimation again does not make it easy to understand what happens here. The main point is simply as expected; one obtains significantly more reproducible/consistent results as N grows.

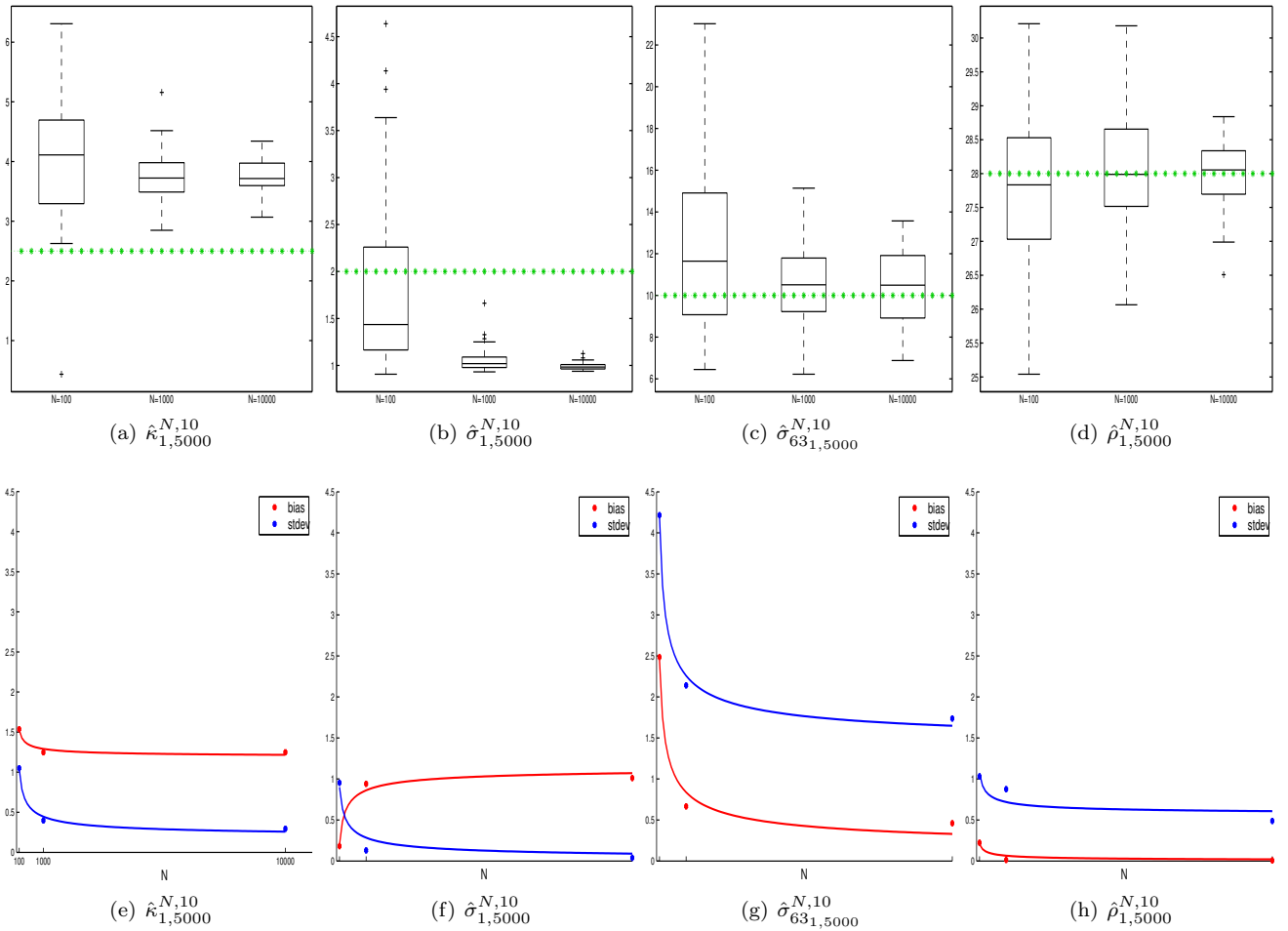


Figure 5: $\hat{\theta}_{1,5000}^{N,10}$ when estimating $\theta = (\kappa, \sigma, \sigma_{63}, \rho)$ of the Lorenz '63 HMM, using ABC-SMC with values of $N \in \{100, 1000, 10000\}$. Figures 5(a)-5(d) show the $\hat{\theta}_{1,5000}^{N,10}$ in boxplots and their true values in dotted green lines. Figures 5(e)-5(h) show the MC bias and MC standard deviation of the $\hat{\theta}_{1,5000}^{N,10}$, in red and blue, with curves of least squared-error $\propto \frac{1}{\sqrt{N}}$.

Next we look at the influence of the pseudo-observations. For $M \in \{1, 3, 5, 10, 25, 50\}$, we show in Figures 6(a)-6(d) the boxplots of the terminal estimates $\hat{\theta}_{1,5000}^{5000,M}$ from fifty independent runs of ABC-SMC, using $N = 5000$ and $\epsilon = 1$. The dotted green lines marks the true θ values which generate the data. In Figures 6(e)-6(h), the MC

biases and the MC standard deviations of the $\hat{\theta}_{1,5000}^{5000,M}$ are plotted point-wise, in red and blue, with lines of least squared-error fit to them. As M increases, we see reductions in the MC variance. This reduction in variance can be attributed to the fact that the ABC-SMC algorithm approximates an algorithm that does not simulate the pseudo data; hence by a Rao-Blackwellization argument, one expects a reduction in variance. These results are consistent with [12]. For this example, after $M \geq 5$, there seems to be little impact on the accuracy of the estimates; it is not clear whether such performance occurs for other examples.

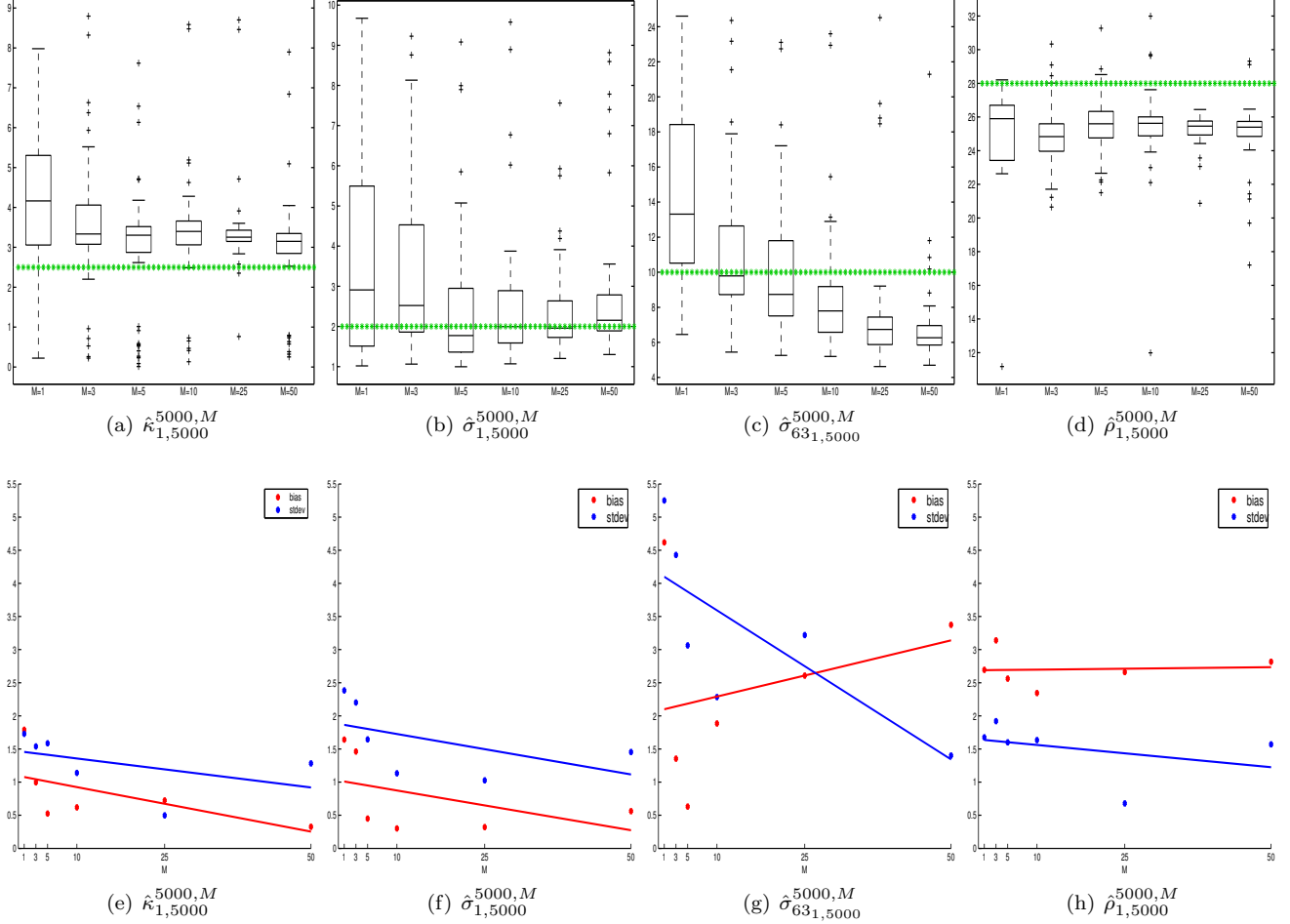


Figure 6: $\hat{\theta}_{1,5000}^{5000,M}$ when estimating $\theta = (\kappa, \sigma, \sigma_{63}, \rho)$ of the Lorenz '63 HMM, using ABC-SMC with values of $M \in \{1, 3, 5, 10, 25, 50\}$. Figures 6(a)-6(d) show the $\hat{\theta}_{1,5000}^{5000,M}$ in boxplots and their true values in dotted green lines. Figures 6(e)-6(h) show the MC bias and MC standard deviation of the $\hat{\theta}_{1,5000}^{5000,M}$, in red and blue, with lines of least squared-error.

We now vary n ; for $n \in \{5000, 10,000, 15,000\}$. We ran fifty independent runs of ABC-SMC using $N = 200$, $M = 10$, and $\epsilon = 1$, and plotted boxplots of the terminal estimates $\hat{\theta}_{1,n}^{200,10}$, in Figures 7(a)-7(d), against the true values of θ marked in dotted green lines. Recall that recursive maximum likelihood estimation tries to maximise $\frac{1}{n} \log(p_{\theta,\epsilon}(y_{1:n}))$, so we expect n not to have a great effect on the bias nor the variance (also due to the bias results in Section 2.3 and the subsequent consistency results in [8, 9]). This is confirmed in Figures 7(e)-7(h), where the absolute value of the MC biases and the MC standard deviations have been plotted in red and blue, and fitted with linear lines of least squared-error.

Finally, we investigate the influence of $\epsilon \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 50\}$. For each ϵ , we again ran fifty independent runs of ABC-SMC with $N = 200$ and $M = 10$, for the dataset $n = 5000$. The boxplot of the parameter estimates are plotted, in Figures 8(a)-8(d), against dotted green lines which indicate the true θ . Figures 8(e)-8(h) show the absolute value of MC biases in red, and the MC standard deviations in blue. Fitted to the MC biases is a non-linear least squares curve proportional to $\epsilon + \frac{1}{\epsilon}$. The result we presented in Section 2.3 states that as ϵ increases, the bias will increase on $\mathcal{O}(\epsilon)$, hence the term proportional to ϵ of the fitted curve. However, the ABC-SMC algorithm

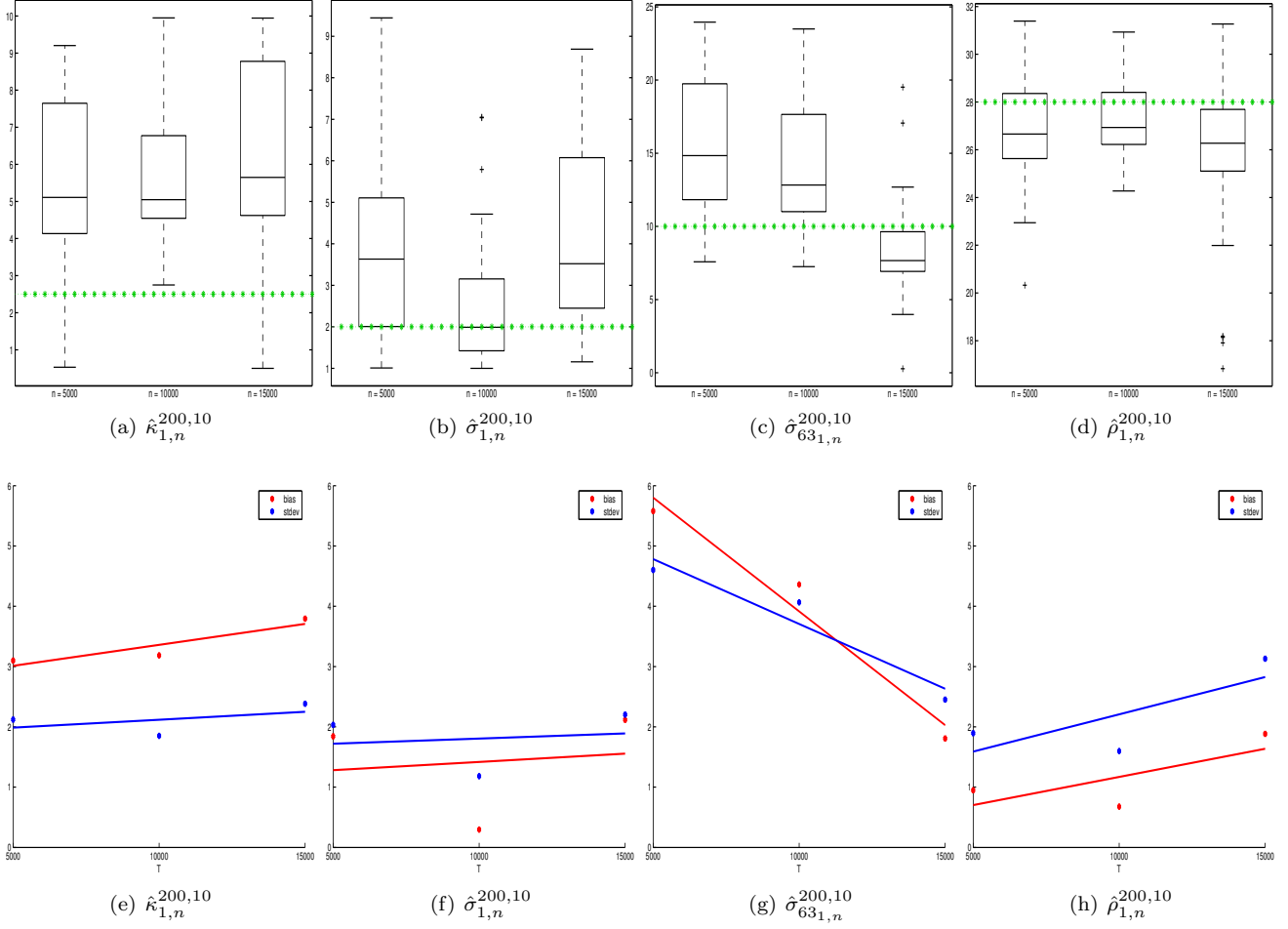


Figure 7: $\hat{\theta}_{1,n}^{200,10}$ when using ABC-SMC to estimate $\theta = (\kappa, \sigma, \sigma_{63}, \rho)$ of the Lorenz '63 HMM, for datasets of length $n \in \{5000, 10000, 15000\}$. Figures 7(a)-7(d) show the $\hat{\theta}_{1,n}^{200,10}$ in boxplots and their true values in dotted green lines. Figures 7(e)-7(h) show the MC bias and MC standard deviation of the $\hat{\theta}_{1,n}^{200,10}$, in red and blue, with lines of least squared-error.

becomes less stable for ϵ too small (in the sense that, for example, the variance of the weights will become larger as ϵ grows), incurring more varied estimates and affected biases; thus the term proportional to $\frac{1}{\epsilon}$. Fitted to the MC standard deviations is a non-linear least squares curves proportional to $\frac{1}{\epsilon}$. For this example, the MC standard deviation decreases at this rate as ϵ increases.

5 Summary

In this article we have presented a technique to perform online parameter estimation using ABC-SMC and SPSA for HMMs. This is useful for models where the state-dimension is high and the parameter and observations are of moderate dimension. In addition, it is required when the conditional density of the observations given the hidden state is intractable.

Some future work is as follows. The representation in Remark 3.1 can be potentially useful for alternative online parameter estimation techniques, other than using SPSA. In [13] we are investigating the use of the online EM algorithm [31] and any potential benefit that it may have over the ideas in this paper. We have remarked that one drawback of the SMC algorithm implemented is its inability to deal with small ϵ . Two potential ways to proceed are as follows. One is to introduce a further approximation by the expectation-propagation algorithm (as in [2]) and potentially removing SMC altogether. The other is to consider more advanced SMC approaches such as [11] and how this might help one reduce ϵ ; this is an area of ongoing research. We are also considering ABC

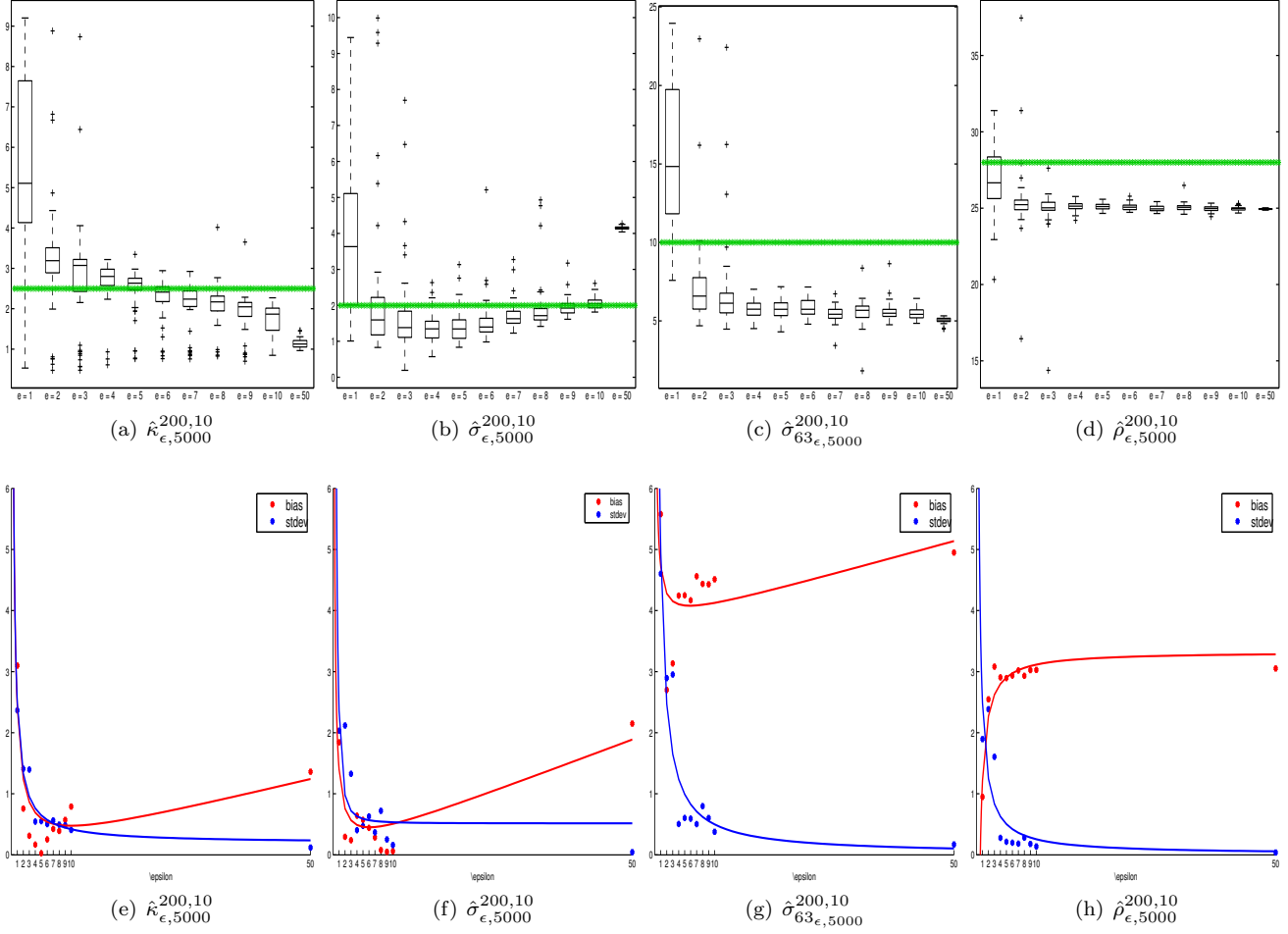


Figure 8: $\hat{\theta}_{\epsilon,5000}^{200,10}$ when estimating $\theta = (\kappa, \sigma, \sigma_{63}, \rho)$ of the Lorenz '63 HMM, using ABC-SMC with values of $\epsilon \in \{1, 2, 3, \dots, 10, 50\}$. Figures 8(a)-8(d) show the MC biases and their curves of non-linear least squared-error proportional to $\epsilon + \frac{1}{\epsilon}$ in red, and the MC standard deviations with their curves of non-linear least squared-error proportional to $\frac{1}{\epsilon}$ in blue.

approximations in the scenario of deterministic dynamics for the hidden state; these models have wide application in applied mathematics as filtering initial conditions of partial differential equations.

Acknowledgements

The second author was funded by an MOE grant and acknowledges useful conversations with David Nott. We also acknowledge useful conversations with Sumeetpal Singh.

A Notations

We introduce a round of notations. As our analysis will rely upon that in [29] our notations will follow that article. It is remarked that under our assumptions, one can establish the same assumptions as in [29]. Moreover, the time-inhomogenous upper-bounds in that paper can be made time-homogenous (albeit less tight) under our assumptions. In addition, our proof strategy follows ideas in [1].

$\mathcal{B}_b(\mathbf{X})$ is the class of bounded and real-valued measurable functions on \mathbf{X} . Throughout, for $\varphi \in \mathcal{B}_b(\mathbf{X})$, $\|\varphi\|_\infty := \sup_{x \in \mathbf{X}} |\varphi(x)|$. For $\varphi \in \mathcal{B}_b(\mathbf{X})$ and any operator $Q : \mathbf{X} \rightarrow \mathcal{M}(\mathbf{X})$, $Q(\varphi)(x) := \int_{\mathbf{X}} \varphi(y) Q(x, dy)$. In addition for $\mu_\theta \in \mathcal{M}(\mathbf{X})$, $\mu_\theta Q(\varphi) := \int_{\mathbf{X}} \mu_\theta(dx) Q(\varphi)(x)$.

We introduce the non-negative operator:

$$R_{\theta,n}(x, dx') := g_\theta(y_n | x') f_\theta(x' | x) dx'$$

with the ABC equivalent $R_{\theta,\epsilon,n}(x, dx') := g_{\theta,\epsilon}(y_n|x')f_{\theta}(x'|x)dx'$, $g_{\theta,\epsilon}(y|x) = \int_{A_{\epsilon,y}} g(u|x)dy / \int_{A_{\epsilon,y}} dy$. To keep consistency with [29] and to allow the reader to follow the proofs, we note that the filter at time $n \geq 0$, $F_{\theta}^n(\mu_{\theta})$ (resp. ABC filter, at time n , $F_{\theta,\epsilon}^n(\mu_{\theta})$) is exactly, with initial distribution $\mu_{\theta} \in \mathcal{P}(\mathbf{X})$ and test function $\varphi \in \mathcal{B}_b(\mathbf{X})$

$$F_{\theta}^n(\mu_{\theta})(\varphi) = \frac{\mu_{\theta}R_{1,n,\theta}(\varphi)}{\mu_{\theta}R_{1,n,\theta}(1)}$$

resp.

$$F_{\theta,\epsilon}^n(\mu_{\theta})(\varphi) = \frac{\mu_{\theta}R_{1,n,\theta,\epsilon}(\varphi)}{\mu_{\theta}R_{1,n,\theta,\epsilon}(1)}$$

where $F_{\theta}^0(\mu_{\theta}) = F_{\theta,\epsilon}^0(\mu_{\theta}) = \mu_{\theta}$, $R_{1,n,\theta}(\varphi)(x_0) = \int \prod_{k=1}^n R_{k,\theta}(x_{k-1}, dx_k)\varphi(x_n)$. In addition, we write the filter derivatives as $\tilde{F}_{\theta}^n(\mu_{\theta}, \tilde{\mu}_{\theta})(\varphi)$, $\tilde{F}_{\theta,\epsilon}^n(\mu_{\theta}, \tilde{\mu}_{\theta})(\varphi)$ where the second argument is the gradient of the initial measure.

The following operators will be used below, for $n \geq 1$:

$$\tilde{G}^n(\mu_{\theta}, \tilde{\mu}_{\theta})(\varphi) := (\mu_{\theta}R_{1,n,\theta}(1))^{-1}[\tilde{\mu}_{\theta}R_{1,n,\theta}(\varphi) - \tilde{\mu}_{\theta}R_{1,n,\theta}(1)F_{\theta}^n(\mu_{\theta})(\varphi)] \quad (10)$$

$$\tilde{H}^n(\mu_{\theta})(\varphi) := F_{\theta}^{n-1}(\mu_{\theta})R_{n,\theta}(1)^{-1}[F_{\theta}^{n-1}(\mu_{\theta})\tilde{R}_{n,\theta}(\varphi) - F_{\theta}^{n-1}(\mu_{\theta})\tilde{R}_{n,\theta}(1)F_{\theta}^n(\mu_{\theta})(\varphi)] \quad (11)$$

with the convention $\tilde{G}^0(\mu_{\theta}, \tilde{\mu}_{\theta})(\varphi) = \tilde{\mu}_{\theta}$. In addition, we set

$$\tilde{G}^{(n)}(\mu_{\theta}, \tilde{\mu}_{\theta})(\varphi) := (\mu_{\theta}R_{n,\theta}(1))^{-1}[\tilde{\mu}_{\theta}R_{n,\theta}(\varphi) - \tilde{\mu}_{\theta}R_{n,\theta}(1)F_{\theta}^{(n)}(\mu_{\theta})(\varphi)].$$

where $F_{\theta}^{(n)}(\mu_{\theta}) = \mu_{\theta}R_{n,\theta}/\mu_{\theta}R_{n,\theta}(1)$. Finally, an important notational convention is as follows. Throughout we use C to denote a constant whose value may change from line-to-line in the calculations. This constant will typically not depend upon important parameters such as ϵ and n and any important dependencies will be highlighted.

B Bias of the Log-Likelihood

Proof of Proposition 2.1. We begin with the equality

$$\log(p_{\theta}(y_{1:n})) - \log(p_{\theta,\epsilon}(y_{1:n})) = \sum_{k=1}^n \left(\log(p_{\theta}(y_k|y_{1:k-1})) - \log(p_{\theta,\epsilon}(y_k|y_{1:k-1})) \right) \quad (12)$$

with, for $1 \leq k \leq n$

$$\begin{aligned} p_{\theta}(y_k|y_{1:k-1}) &= \int_{\mathbf{X}^2} g_{\theta}(y_k|x_k)f_{\theta}(x_k|x_{k-1})F_{\theta}^{k-1}(\mu_{\theta})(dx_{k-1})dx_k \\ p_{\theta,\epsilon}(y_k|y_{1:k-1}) &= \int_{\mathbf{X}^2} g_{\theta,\epsilon}^{\epsilon}(y_k|x_k)f_{\theta}(x_k|x_{k-1})F_{\theta,\epsilon}^{k-1}(\mu_{\theta})(dx_{k-1})dx_k. \end{aligned}$$

We will consider each summand in (12). The case $k \geq 2$ is only considered; the scenario $k = 1$ will follow a similar and simpler argument.

Using the inequality $|\log(x) - \log(y)| \leq |x - y|/(x \wedge y)$ for every $x, y > 0$ we have

$$|\log(p_{\theta}(y_k|y_{1:k-1})) - \log(p_{\theta,\epsilon}(y_k|y_{1:k-1}))| \leq \frac{|p_{\theta}(y_k|y_{1:k-1}) - p_{\theta,\epsilon}(y_k|y_{1:k-1})|}{p_{\theta}(y_k|y_{1:k-1}) \wedge p_{\theta,\epsilon}(y_k|y_{1:k-1})}.$$

Note that

$$\begin{aligned} p_{\theta}(y_k|y_{1:k-1}) \wedge p_{\theta,\epsilon}(y_k|y_{1:k-1}) &= \\ \int_{\mathbf{X}^2} g_{\theta}(y_k|x_k)f_{\theta}(x_k|x_{k-1})F_{\theta}^{k-1}(\mu_{\theta})(dx_{k-1})dx_k \wedge \int_{\mathbf{X}^2} g_{\theta,\epsilon}^{\epsilon}(y_k|x_k)f_{\theta}(x_k|x_{k-1})F_{\theta,\epsilon}^{k-1}(\mu_{\theta})(dx_{k-1})dx_k &\geq C > 0 \end{aligned} \quad (13)$$

where we have applied (A3) and C does not depend upon ϵ . Thus we consider

$$\begin{aligned} |p_{\theta,\epsilon}(y_k|y_{1:k-1}) - p_{\theta}(y_k|y_{1:k-1})| &= \\ \left| \int_{\mathbf{X}^2} g_{\theta}(y_k|x_k)f_{\theta}(x_k|x_{k-1})F_{\theta}^{k-1}(\mu_{\theta})(dx_{k-1})dx_k - \int_{\mathbf{X}^2} g_{\theta,\epsilon}^{\epsilon}(y_k|x_k)f_{\theta}(x_k|x_{k-1})F_{\theta,\epsilon}^{k-1}(\mu_{\theta})(dx_{k-1})dx_k \right|. \end{aligned}$$

The R.H.S. can be upper-bounded by the sum of

$$| \int_{\mathcal{X}^2} [g_\theta(y_k|x_k) - g_{\theta,\epsilon}(y_k|x_k)] f_\theta(x_k|x_{k-1}) F_\theta^{k-1}(\mu_\theta)(dx_{k-1}) dx_k |$$

and

$$| \int_{\mathcal{X}^2} g_{\theta,\epsilon}(y_k|x_k) f_\theta(x_k|x_{k-1}) [F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1}) - F_\theta^{k-1}(\mu_\theta)(dx_{k-1})] dx_k |.$$

The first expression can be dealt with by using (A1), which implies

$$\sup_{x \in \mathcal{X}} |g_{\theta,\epsilon}(y_k|x) - g_\theta(y_k|x)| \leq C\epsilon. \quad (14)$$

The second expression can be controlled by [18, Theorem 2]:

$$\sup_{k \geq 1} \|F_\theta^{k-1}(\mu_\theta) - F_{\theta,\epsilon}^{k-1}(\mu_\theta)\| \leq C\epsilon \quad (15)$$

to yield that

$$|p_{\theta,\epsilon}(y_k|y_{1:k-1}) - p_\theta(y_k|y_{1:k-1})| \leq C\epsilon. \quad (16)$$

One can thus conclude. \square

C Bias of the Gradient of the Log-Likelihood

Proof of Theorem 2.1. We have that

$$\nabla \left(\log p_\theta(y_{1:n}) - \log p_{\theta,\epsilon}(y_{1:n}) \right) = \nabla \left\{ \sum_{k=1}^n \left(\log[p_\theta(y_k|y_{1:k-1})] - \log[p_{\theta,\epsilon}(y_k|y_{1:k-1})] \right) \right\}.$$

It then follows that

$$\begin{aligned} & \nabla \left(\log p_\theta(y_{1:n}) - \log p_{\theta,\epsilon}(y_{1:n}) \right) = \\ & \sum_{k=1}^n \left(\frac{[\nabla p_\theta(y_k|y_{1:k-1}) - \nabla p_{\theta,\epsilon}(y_k|y_{1:k-1})]}{p_\theta(y_k|y_{1:k-1})} + \frac{\nabla p_{\theta,\epsilon}(y_k|y_{1:k-1})}{p_\theta(y_k|y_{1:k-1}) p_{\theta,\epsilon}(y_k|y_{1:k-1})} [p_{\theta,\epsilon}(y_k|y_{1:k-1}) - p_\theta(y_k|y_{1:k-1})] \right). \end{aligned} \quad (17)$$

We will deal with the two terms on the R.H.S. of (17) in turn. The scenario $k \geq 2$ is only considered; the case $k = 1$ follows a similar and simpler argument.

First starting with summand

$$\frac{[\nabla p_\theta(y_k|y_{1:k-1}) - \nabla p_{\theta,\epsilon}(y_k|y_{1:k-1})]}{p_\theta(y_k|y_{1:k-1})}.$$

Noting (13), we need only upper-bound the \mathbb{L}_1 norm of the following expression

$$\int_{\mathcal{X}^2} \nabla \{g_\theta(y_k|x_k)\} f_\theta(x_k|x_{k-1}) F_\theta^{k-1}(\mu_\theta)(dx_{k-1}) dx_k - \int_{\mathcal{X}^2} \nabla \{g_{\theta,\epsilon}(y_k|x_k)\} f_\theta(x_k|x_{k-1}) F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1}) dx_k \quad (18)$$

$$+ \int_{\mathcal{X}^2} g_\theta(y_k|x_k) \nabla \{f_\theta(x_k|x_{k-1})\} F_\theta^{k-1}(\mu_\theta)(dx_{k-1}) dx_k - \int_{\mathcal{X}^2} g_{\theta,\epsilon}(y_k|x_k) \nabla \{f_\theta(x_k|x_{k-1})\} F_{\theta,\epsilon}^{k-1}(\mu_\theta)(dx_{k-1}) dx_k \quad (19)$$

$$+ \int_{\mathcal{X}^2} g_\theta(y_k|x_k) f_\theta(x_k|x_{k-1}) \tilde{F}_\theta^{k-1}(\mu_\theta, \tilde{\mu}_\theta)(dx_{k-1}) dx_k - \int_{\mathcal{X}^2} g_{\theta,\epsilon}(y_k|x_k) f_\theta(x_k|x_{k-1}) \tilde{F}_{\theta,\epsilon}^{k-1}(\mu_\theta, \tilde{\mu}_\theta)(dx_{k-1}) dx_k. \quad (20)$$

We start with (18). Using (A4) we can establish that for each $k \geq 1$

$$\sup_{x \in \mathcal{X}} |\nabla \{g_\theta(y_k|x_k)\} - \nabla \{g_{\theta,\epsilon}(y_k|x_k)\}| \leq C\epsilon \quad (21)$$

where C does not depend upon k, ϵ . Hence

$$| \int_{\mathcal{X}^2} [\nabla \{g_\theta(y_k|x_k)\} - \nabla \{g_{\theta,\epsilon}(y_k|x_k)\}] f_\theta(x_k|x_{k-1}) F_\theta^{k-1}(\mu_\theta)(dx_{k-1}) dx_k | \leq C\epsilon.$$

Then we note that by [18, Theorem 2] (see (15)) and (A5)

$$| \int_{\mathcal{X}^2} \nabla \{g_{\theta, \epsilon}(y_k | x_k)\} f_{\theta}(x_k | x_{k-1}) [F_{\theta}^{k-1}(\mu_{\theta})(dx_{k-1}) - F_{\theta, \epsilon}^{k-1}(\mu_{\theta})(dx_{k-1})] dx_k | \leq C\epsilon$$

Thus we have shown that

$$| \int_{\mathcal{X}^2} \nabla \{g_{\theta}(y_k | x_k)\} f_{\theta}(x_k | x_{k-1}) F_{\theta}^{k-1}(\mu_{\theta})(dx_{k-1}) dx_k - \int_{\mathcal{X}^2} \nabla \{g_{\theta, \epsilon}(y_k | x_k)\} f_{\theta}(x_k | x_{k-1}) F_{\theta, \epsilon}^{k-1}(\mu_{\theta})(dx_{k-1}) dx_k | \leq C\epsilon.$$

Now, moving onto (19), by (14) we have

$$| \int_{\mathcal{X}^2} [g_{\theta}(y_k | x_k) - g_{\theta, \epsilon}(y_k | x_k)] \nabla \{f_{\theta}(x_k | x_{k-1})\} F_{\theta}^{k-1}(\mu_{\theta})(dx_{k-1}) dx_k | \leq C\epsilon.$$

and can again use [18, Theorem 2] (i.e. (15)) to deduce that

$$| \int_{\mathcal{X}^2} g_{\theta, \epsilon}(y_k | x_k) \nabla \{f_{\theta}(x_k | x_{k-1})\} [F_{\theta}^{k-1}(\mu_{\theta})(dx_{k-1}) - F_{\theta, \epsilon}^{k-1}(\mu_{\theta})(dx_{k-1})] dx_k | \leq C\epsilon$$

and thus that

$$| \int_{\mathcal{X}^2} g_{\theta}(y_k | x_k) \nabla \{f_{\theta}(x_k | x_{k-1})\} F_{\theta}^{k-1}(\mu_{\theta})(dx_{k-1}) dx_k - \int_{\mathcal{X}^2} g_{\theta, \epsilon}(y_k | x_k) \nabla \{f_{\theta}(x_k | x_{k-1})\} F_{\theta, \epsilon}^{k-1}(\mu_{\theta})(dx_{k-1}) dx_k | \leq C\epsilon$$

which upper-bounds the expression in (19). We now move onto (20), which upper-bounded by

$$\begin{aligned} & | \int_{\mathcal{X}^2} [g_{\theta}(y_k | x_k) - g_{\theta, \epsilon}(y_k | x_k)] f_{\theta}(x_k | x_{k-1}) \tilde{F}_{\theta}^{k-1}(\mu_{\theta}, \widetilde{\mu}_{\theta})(dx_{k-1}) dx_k | + \\ & | \int_{\mathcal{X}^2} g_{\theta, \epsilon}(y_k | x_k) f_{\theta}(x_k | x_{k-1}) [\tilde{F}_{\theta}^{k-1}(\mu_{\theta}, \widetilde{\mu}_{\theta})(dx_{k-1}) - \tilde{F}_{\theta, \epsilon}^{k-1}(\mu_{\theta}, \widetilde{\mu}_{\theta})(dx_{k-1})] dx_k |. \end{aligned}$$

For the first expression, we can write:

$$(\sup_{x \in \mathcal{X}} |g_{\theta}(y_k | x) - g_{\theta, \epsilon}(y_k | x)|) | \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \frac{[g_{\theta}(y_k | x_k) - g_{\theta, \epsilon}(y_k | x_k)]}{(\sup_{x \in \mathcal{X}} |g_{\theta}(y_k | x) - g_{\theta, \epsilon}(y_k | x)|)} f_{\theta}(x_k | x_{k-1}) dx_k \right) \tilde{F}_{\theta}^{k-1}(\mu_{\theta}, \widetilde{\mu}_{\theta})(dx_{k-1}) |.$$

Then we can apply (14) and, noting that

$$\left(\int_{\mathcal{X}} \frac{[g_{\theta}(y_k | x_k) - g_{\theta, \epsilon}(y_k | x_k)]}{(\sup_{x \in \mathcal{X}} |g_{\theta}(y_k | x) - g_{\theta, \epsilon}(y_k | x)|)} f_{\theta}(x_k | x_{k-1}) dx_k \right) \leq 1$$

one can also use Lemma D.3 to deduce that

$$| \int_{\mathcal{X}^2} [g_{\theta}(y_k | x_k) - g_{\theta, \epsilon}(y_k | x_k)] f_{\theta}(x_k | x_{k-1}) \tilde{F}_{\theta}^{k-1}(\mu_{\theta}, \widetilde{\mu}_{\theta})(dx_{k-1}) dx_k | \leq C(1 + \|\widetilde{\mu}_{\theta}\|)\epsilon.$$

Then, one can easily apply Theorem D.1 to show that

$$| \int_{\mathcal{X}^2} g_{\theta, \epsilon}(y_k | x_k) f_{\theta}(x_k | x_{k-1}) [\tilde{F}_{\theta}^{k-1}(\mu_{\theta}, \widetilde{\mu}_{\theta})(dx_{k-1}) - \tilde{F}_{\theta, \epsilon}^{k-1}(\mu_{\theta}, \widetilde{\mu}_{\theta})(dx_{k-1})] dx_k | \leq C(2 + \|\widetilde{\mu}_{\theta}\|)\epsilon.$$

Thus we have upper-bounded the \mathbb{L}_1 -norm of the sum of the expressions (18)-(20) and we have established that

$$\frac{[\nabla p_{\theta}(y_k | y_{1:k-1}) - \nabla p_{\theta, \epsilon}(y_k | y_{1:k-1})]}{p_{\theta}(y_k | y_{1:k-1})} \leq C(2 + \|\widetilde{\mu}_{\theta}\|)\epsilon. \quad (22)$$

Moving onto the second summand on the R.H.S. of (17),

$$\frac{\nabla p_{\theta, \epsilon}(y_k | y_{1:k-1})}{p_{\theta}(y_k | y_{1:k-1}) p_{\theta, \epsilon}(y_k | y_{1:k-1})} [p_{\theta, \epsilon}(y_k | y_{1:k-1}) - p_{\theta}(y_k | y_{1:k-1})].$$

By (16), we need only consider upper-bounding, in \mathbb{L}_1 , $\nabla p_{\theta,\epsilon}(y_k|y_{1:k-1})$. This can be decomposed into the sum of three expressions:

$$\begin{aligned} & \int_{\mathbf{X}^2} \nabla \{g_{\theta,\epsilon}(y_k|x_k)\} f_{\theta}(x_k|x_{k-1}) F_{\theta,\epsilon}^{k-1}(\mu_{\theta})(dx_{k-1}) dx_k \\ & \int_{\mathbf{X}^2} g_{\theta,\epsilon}(y_k|x_k) \nabla \{f_{\theta}(x_k|x_{k-1})\} F_{\theta,\epsilon}^{k-1}(\mu_{\theta})(dx_{k-1}) dx_k \end{aligned}$$

and

$$\int_{\mathbf{X}^2} g_{\theta,\epsilon}(y_k|x_k) f_{\theta}(x_k|x_{k-1}) \widetilde{F}_{\theta,\epsilon}^{k-1}(\mu_{\theta}, \widetilde{\mu}_{\theta})(dx_{k-1}) dx_k.$$

As $\nabla \{g_{\theta,\epsilon}(y_k|x_k)\}$ and $g_{\theta,\epsilon}(y_k|x_k) \nabla \{f_{\theta}(x_k|x_{k-1})\}$ are upper-bounded as well as \mathbf{X} being compact the first two expressions are upper-bounded in \mathbb{L}_1 . In addition as $\int_{\mathbf{X}} g_{\theta,\epsilon}(y_k|x_k) f_{\theta}(x_k|x_{k-1}) dx_k$ is upper-bounded, we can apply Lemma D.3 to see that the third expression is upper-bounded in \mathbb{L}_1 . Hence, we have shown that

$$\left| \frac{\nabla p_{\theta,\epsilon}(y_k|y_{1:k-1})}{p_{\theta}(y_k|y_{1:k-1}) p_{\theta,\epsilon}(y_k|y_{1:k-1})} [p_{\theta,\epsilon}(y_k|y_{1:k-1}) - p_{\theta}(y_k|y_{1:k-1})] \right| \leq C(1 + \|\widetilde{\mu}_{\theta}\|)\epsilon. \quad (23)$$

Combining the results (22)-(23) and noting (17) we can conclude. \square

D Bias of the Gradient of the Filter

Theorem D.1. *Assume (A1-5). Then there exist a $C < +\infty$ such that for any $n \geq 1$, $\mu_{\theta} \in \mathcal{P}(\mathbf{X})$, $\widetilde{\mu}_{\theta} \in \mathcal{M}(\mathbf{X})$, $\epsilon > 0$, $\theta \in \Theta$:*

$$\|\widetilde{F}_{\theta}^n(\mu_{\theta}, \widetilde{\mu}_{\theta}) - \widetilde{F}_{\theta,\epsilon}^n(\mu_{\theta}, \widetilde{\mu}_{\theta})\| \leq C\epsilon(2 + \|\widetilde{\mu}_{\theta}\|).$$

Proof. We have the following telescoping sum decomposition (e.g. [10]) for the differences in the filters, with $\varphi \in \mathcal{B}_b(\mathbf{X})$:

$$F_{\theta}^n(\mu_{\theta})(\varphi) - F_{\theta,\epsilon}^n(\mu_{\theta})(\varphi) = \sum_{p=1}^n \left[F_{\theta}^{n-p+1,n}(F_{\theta,\epsilon}^{n-p}(\mu_{\theta}))(\varphi) - F_{\theta}^{n-p+2,n}(F_{\theta,\epsilon}^{n-p+1}(\mu_{\theta}))(\varphi) \right]$$

where we are using the notation $F_{\theta}^{q,n}(\mu_{\theta})(\varphi) = \frac{\mu_{\theta} R_{q,n,\theta}(\varphi)}{\mu_{\theta} R_{q,n,\theta}(1)}$, for $1 \leq q \leq n$. Hence, taking gradients and swapping the order of summation and differentiation we have and omitting the second arguments of \widetilde{F} on the R.H.S. (to reduce the notational burden)

$$\begin{aligned} \widetilde{F}_{\theta}^n(\mu_{\theta}, \widetilde{\mu}_{\theta})(\varphi) - \widetilde{F}_{\theta,\epsilon}^n(\mu_{\theta}, \widetilde{\mu}_{\theta})(\varphi) &= \sum_{p=1}^n \left[\widetilde{F}_{\theta}^{n-p+2,n}(F_{\theta}^{(n-p+1)}[F_{\theta,\epsilon}^{n-p}(\mu_{\theta})], \widetilde{F}_{\theta}^{(n-p+1)}[F_{\theta,\epsilon}^{n-p}(\mu_{\theta})])(\varphi) - \right. \\ &\quad \left. \widetilde{F}_{\theta}^{n-p+2,n}(F_{\theta,\epsilon}^{(n-p+1)}[F_{\theta,\epsilon}^{(n-p)}(\mu_{\theta})], \widetilde{F}_{\theta,\epsilon}^{(n-p+1)}[F_{\theta,\epsilon}^{(n-p)}(\mu_{\theta})])(\varphi) \right]. \quad (24) \end{aligned}$$

To continue with the proof we will adopt [29, Lemma 6.4]:

$$\widetilde{F}_{\theta}^n(\mu_{\theta}, \widetilde{\mu}_{\theta})(\varphi) = \widetilde{G}_{\theta}^n(\mu_{\theta}, \widetilde{\mu}_{\theta}) + \sum_{q=1}^n \widetilde{G}_{\theta}^{q+1,n}(F_{\theta}^q(\mu_{\theta}), \widetilde{H}^q(\mu_{\theta}))(\varphi)$$

with \widetilde{G}_{θ}^n and $\widetilde{H}^q(\mu_{\theta})$ defined in (10)-(11) and $\widetilde{G}_{\theta}^{q+1,n}$ similar extension to the notation as for the filter $F_{\theta}^{q+1,n}$ and the convention $\widetilde{G}_{\theta}^{n+1,n}(\mu_{\theta}, \widetilde{\mu}_{\theta}) = \widetilde{\mu}_{\theta}$. Returning to (24) and again omitting the second arguments of \widetilde{F} on the R.H.S.:

$$\begin{aligned} & \widetilde{F}_{\theta}^n(\mu_{\theta}, \widetilde{\mu}_{\theta})(\varphi) - \widetilde{F}_{\theta,\epsilon}^n(\mu_{\theta}, \widetilde{\mu}_{\theta})(\varphi) = \\ & \sum_{p=1}^n \left[\widetilde{G}_{\theta}^{n-p+2,n} \{ F_{\theta}^{(n-p+1)}(F_{\theta,\epsilon}^{n-p}(\mu_{\theta})), \widetilde{F}_{\theta}^{(n-p+1)}(F_{\theta,\epsilon}^{n-p}(\mu_{\theta})) \}(\varphi) - \widetilde{G}_{\theta}^{n-p+2,n} \{ F_{\theta,\epsilon}^{(n-p+1)}[F_{\theta,\epsilon}^{n-p}(\mu_{\theta})], \widetilde{F}_{\theta,\epsilon}^{(n-p+1)}[F_{\theta,\epsilon}^{n-p}(\mu_{\theta})] \}(\varphi) + \right. \\ & \quad \left. \sum_{q=n-p+2}^n \left\{ \widetilde{G}_{\theta}^{q+1,n} \{ F_{\theta}^{n-p+2,q}[F_{\theta}^{(n-p+1)}(F_{\theta,\epsilon}^{n-p}(\mu_{\theta}))], \widetilde{H}_{\theta}^{n-p+2,q}[F_{\theta}^{(n-p+1)}(F_{\theta,\epsilon}^{n-p}(\mu_{\theta}))] \}(\varphi) - \right. \right. \end{aligned}$$

$$\left. \tilde{G}_\theta^{q+1,n} \{F_\theta^{n-p+2,q}[F_{\theta,\epsilon}^{(n-p+1)}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^{n-p+2,q}[F_{\theta,\epsilon}^{(n-p+1)}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) \right\}. \quad (25)$$

We start first with the summand on the R.H.S. of the second line of (25), which we compactly denote as:

$$\tilde{G}_\theta^{p-1} \{F_\theta[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \tilde{G}_\theta^{p-1} \{F_{\theta,\epsilon}[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi).$$

This can be decomposed further into the sum of

$$\tilde{G}_\theta^{p-1} \{F_\theta[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \tilde{G}_\theta^{p-1} \{F_{\theta,\epsilon}[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) \quad (26)$$

and

$$\tilde{G}_\theta^{p-1} \{F_{\theta,\epsilon}[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \tilde{G}_\theta^{p-1} \{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \tilde{F}_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi). \quad (27)$$

Beginning with (26), by [29, Lemma 6.7], equation (43) we have

$$\begin{aligned} & |\tilde{G}_\theta^{p-1} \{F_\theta[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \tilde{G}_\theta^{p-1} \{F_{\theta,\epsilon}[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi)| \\ & \leq C \|\varphi\|_\infty \rho^{p-1} \|F_\theta[F_\theta^{n-p}(\mu_\theta)] - F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\| \|\tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\| \end{aligned}$$

where $\rho \in (0, 1)$ and C do not depend upon μ_θ, ϵ or n, p . Applying Lemma D.2 we have

$$\begin{aligned} & |\tilde{G}_\theta^{p-1} \{F_\theta[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \tilde{G}_\theta^{p-1} \{F_{\theta,\epsilon}[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi)| \\ & \leq C \|\varphi\|_\infty \rho^{p-1} \epsilon \|\tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\| \end{aligned}$$

where C does not depend upon μ_θ, ϵ or n, p . Then by Remark D.1 and Lemma D.3 $\|\tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\| \leq C(2 + \|\tilde{\mu}_\theta\|)$ and thus the upper-bound on the \mathbb{L}_1 -norm of (26):

$$|\tilde{G}_\theta^{p-1} \{F_\theta[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \tilde{G}_\theta^{p-1} \{F_{\theta,\epsilon}[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi)| \leq C \|\varphi\|_\infty \epsilon \rho^{p-1} (2 + \|\tilde{\mu}_\theta\|). \quad (28)$$

Now, moving onto (27), by [29, Lemma 6.7], equation (42):

$$\begin{aligned} & |\tilde{G}_\theta^{p-1} \{F_{\theta,\epsilon}[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \tilde{G}_\theta^{p-1} \{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \tilde{F}_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi)| \\ & \leq C \rho^{p-1} \|\varphi\|_\infty \|\tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)] - \tilde{F}_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\|. \end{aligned}$$

Applying Lemma D.1

$$\begin{aligned} & |\tilde{G}_\theta^{p-1} \{F_{\theta,\epsilon}[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \tilde{G}_\theta^{p-1} \{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \tilde{F}_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi)| \\ & \leq C \|\varphi\|_\infty \epsilon \rho^{p-1} (1 + \|\tilde{F}_{\theta,\epsilon}^{n-p}(\mu_\theta)\|). \end{aligned}$$

Then by Lemma D.3, we deduce that

$$|\tilde{G}_\theta^{p-1} \{F_{\theta,\epsilon}[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \tilde{G}_\theta^{p-1} \{F_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)], \tilde{F}_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi)| \leq C \|\varphi\|_\infty \epsilon \rho^{p-1} (2 + \|\tilde{\mu}_\theta\|). \quad (29)$$

Combining (28) and (29)

$$|\tilde{G}_\theta^{p-1} \{F_\theta[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_\theta[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi) - \tilde{G}_\theta^{p-1} \{F_{\theta,\epsilon}[F_\theta^{n-p}(\mu_\theta)], \tilde{F}_{\theta,\epsilon}[F_{\theta,\epsilon}^{n-p}(\mu_\theta)]\}(\varphi)| \leq C \|\varphi\|_\infty \epsilon \rho^{p-1} (2 + \|\tilde{\mu}_\theta\|). \quad (30)$$

We now consider the summands over q in the second and third lines of (25). Again, adopting the compact notation above we can decompose the summands over q into the sum of

$$\tilde{G}_\theta^{n-q} \{F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \tilde{G}_\theta^{n-q} \{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) \quad (31)$$

and

$$\tilde{G}_\theta^{n-q} \{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \tilde{G}_\theta^{n-q} \{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) \quad (32)$$

where $s = q - n + p - 1$. We start with (31); by [29, Lemma 6.7] equation (43), we have

$$|\tilde{G}_\theta^{n-q} \{F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \tilde{G}_\theta^{n-q} \{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi)|$$

$$\leq C\|\varphi\|_\infty \rho^{n-q} \|F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))] - F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\| \|\tilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\|.$$

Then we will use the stability of the filter (e.g. [29, Theorem 3.1])

$$\|F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))] - F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\| \leq C\rho^s \|F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta)) - F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))\|.$$

By Lemma D.2 $\|F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta)) - F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))\| \leq C\epsilon$ and thus

$$\begin{aligned} & |\tilde{G}_\theta^{n-q}\{F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \tilde{G}_\theta^{n-q}\{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi)| \\ & \leq C\|\varphi\|_\infty \epsilon \rho^{p-1} \|\tilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\|. \end{aligned}$$

By [29, Lemma 6.8] we have $\|\tilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\| \leq C$, where C does not depend upon $F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))$ or ϵ and hence

$$|\tilde{G}_\theta^{n-q}\{F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \tilde{G}_\theta^{n-q}\{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi)| \leq C\|\varphi\|_\infty \epsilon \rho^{p-1}.$$

Now, turning to (32) and applying [29, Lemma 6.7] (42) we have

$$\begin{aligned} & |\tilde{G}_\theta^{n-q}\{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \tilde{G}_\theta^{n-q}\{F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi)| \\ & \leq C\|\varphi\|_\infty \rho^{n-q} \|\tilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\| - \tilde{H}_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\|. \end{aligned} \quad (33)$$

Then by [29, Lemma 6.8] we have

$$\|\tilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\| - \tilde{H}_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\| \leq C\rho^s \|F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta)) - F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))\|$$

and then on applying Lemma D.2 we thus have that

$$\|\tilde{H}_\theta^q(F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta)) - \tilde{H}_\theta^q(F_{\theta,\epsilon}^{n-p+1}(\mu_\theta))\| \leq C\epsilon\rho^s.$$

Returning to (33), it follows by the above calculations that:

$$|\tilde{G}_\theta^{n-q}\{F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \tilde{G}_\theta^{n-q}\{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi)| \leq C\|\varphi\|_\infty \epsilon \rho^{p-1}.$$

Thus we have proved that

$$|\tilde{G}_\theta^{n-q}\{F_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_\theta(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi) - \tilde{G}_\theta^{n-q}\{F_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))], \tilde{H}_\theta^s[F_{\theta,\epsilon}(F_{\theta,\epsilon}^{n-p}(\mu_\theta))]\}(\varphi)| \leq C\|\varphi\|_\infty \epsilon \rho^{p-1}. \quad (34)$$

Then, returning to (25) and noting (30), (34) we have the upper-bound

$$\|\tilde{F}_\theta^n(\mu_\theta, \tilde{\mu}_\theta) - \tilde{F}_{\theta,\epsilon}^n(\mu_\theta, \tilde{\mu}_\theta)\| \leq C\epsilon(2 + \|\tilde{\mu}_\theta\|) \sum_{p=1}^n [\rho^{p-1} + \sum_{q=n-p}^n \rho^{p-1}] \leq C\epsilon(2 + \|\tilde{\mu}_\theta\|).$$

□

D.1 Technical Results for ABC Bias of the Filter-Derivative

Lemma D.1. *Assume (A1-5). Then there exist a $C < +\infty$ such that for any $n \geq 1$, $\mu_\theta \in \mathcal{P}(\mathbf{X})$, $\tilde{\mu}_\theta \in \mathcal{M}(\mathbf{X})$, $\epsilon > 0$ $\theta \in \Theta$:*

$$\|\tilde{F}_\theta^{(n)}(\mu_\theta, \tilde{\mu}_\theta) - \tilde{F}_{\theta,\epsilon}^{(n)}(\mu_\theta, \tilde{\mu}_\theta)\| \leq C\epsilon(1 + \|\tilde{\mu}_\theta\|).$$

Proof. By [29, Lemma 6.7] we have the decomposition, for $\varphi \in \mathcal{B}_b(\mathbf{X})$:

$$\tilde{F}_\theta^{(n)}(\mu_\theta, \tilde{\mu}_\theta)(\varphi) = \tilde{G}_\theta^{(n)}(\mu_\theta, \tilde{\mu}_\theta)(\varphi) - \tilde{H}_\theta^{(n)}(\mu_\theta)(\varphi)$$

where

$$\tilde{H}_\theta^{(n)}(\mu_\theta)(\varphi) := \mu_\theta R_{n,\theta}(1)^{-1} [\mu_\theta \tilde{R}_{n,\theta}(\varphi) - \mu_\theta \tilde{R}_{n,\theta}(1) \mu_\theta(\varphi)].$$

Thus to control the difference, we can consider the two differences $\tilde{G}_\theta^{(n)}(\mu_\theta, \tilde{\mu}_\theta)(\varphi) - \tilde{G}_{\theta,\epsilon}^{(n)}(\mu_\theta, \tilde{\mu}_\theta)(\varphi)$ and $\tilde{H}_\theta^{(n)}(\mu_\theta)(\varphi) - \tilde{H}_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)$.

Control of $\tilde{G}_\theta^{(n)}(\mu_\theta, \tilde{\mu}_\theta)(\varphi) - \tilde{G}_{\theta,\epsilon}^{(n)}(\mu_\theta, \tilde{\mu}_\theta)(\varphi)$. We will use the Hahn-Jordan decomposition: $\tilde{\mu}_\theta = \tilde{\mu}_\theta^+ - \tilde{\mu}_\theta^-$. It is assumed that both $\tilde{\mu}_\theta^+(1), \tilde{\mu}_\theta^-(1) > 0$. The scenario with either $\tilde{\mu}_\theta^+(1) = 0$ or $\tilde{\mu}_\theta^-(1) = 0$ is straightforward and omitted for brevity. We can write:

$$\tilde{G}_\theta^{(n)}(\mu_\theta, \tilde{\mu}_\theta)(\varphi) = \frac{\tilde{\mu}_\theta^+ R_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)} [F_\theta^{(n)}(\tilde{\mu}_\theta^+)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)] + \frac{\tilde{\mu}_\theta^- R_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)} [F_\theta^{(n)}(\tilde{\mu}_\theta^-)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)]$$

where $\tilde{\mu}_\theta^+(\cdot) = \tilde{\mu}_\theta^+(\cdot)/\tilde{\mu}_\theta^+(1)$ and $\tilde{\mu}_\theta^-(\cdot) = \tilde{\mu}_\theta^-(\cdot)/\tilde{\mu}_\theta^-(1)$. Thus we have

$$\begin{aligned} \tilde{G}_\theta^{(n)}(\mu_\theta, \tilde{\mu}_\theta)(\varphi) - \tilde{G}_{\theta,\epsilon}^{(n)}(\mu_\theta, \tilde{\mu}_\theta)(\varphi) &= \left[\frac{\tilde{\mu}_\theta^+ R_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)} - \frac{\tilde{\mu}_\theta^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)} \right] [F_\theta^{(n)}(\tilde{\mu}_\theta^+)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)] \\ &+ \frac{\tilde{\mu}_\theta^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)} [F_\theta^{(n)}(\tilde{\mu}_\theta^+)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi) - F_{\theta,\epsilon}^{(n)}(\tilde{\mu}_\theta^+)(\varphi) + F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)] \\ &+ \left[\frac{\tilde{\mu}_\theta^- R_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)} - \frac{\tilde{\mu}_\theta^- R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)} \right] [F_\theta^{(n)}(\tilde{\mu}_\theta^-)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)] \\ &+ \frac{\tilde{\mu}_\theta^- R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)} [F_\theta^{(n)}(\tilde{\mu}_\theta^-)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi) - F_{\theta,\epsilon}^{(n)}(\tilde{\mu}_\theta^-)(\varphi) + F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)] \end{aligned} \quad (35)$$

By symmetry, we need only consider the terms including $\tilde{\mu}_\theta^+$; one can treat those with $\tilde{\mu}_\theta^-$ by using similar arguments. First dealing with term on the first line of the R.H.S. of (35). We have that

$$\begin{aligned} &\left[\frac{\tilde{\mu}_\theta^+ R_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)} - \frac{\tilde{\mu}_\theta^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)} \right] [F_\theta^{(n)}(\tilde{\mu}_\theta^+)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)] = \\ &\left[\frac{\tilde{\mu}_\theta^+ R_{n,\theta}(1) - \tilde{\mu}_\theta^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta}(1)} + \tilde{\mu}_\theta^+ R_{n,\theta,\epsilon}(1) \frac{\mu_\theta R_{n,\theta,\epsilon}(1) - \mu_\theta R_{n,\theta}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1) \mu_\theta R_{n,\theta}(1)} \right] [F_\theta^{(n)}(\tilde{\mu}_\theta^+)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)] \end{aligned}$$

Now by (A1), for any n

$$\sup_{x \in \mathbf{X}} |R_{n,\theta}(1)(x) - R_{n,\theta,\epsilon}(1)(x)| \leq C\epsilon \quad (36)$$

thus

$$\left[\frac{\tilde{\mu}_\theta^+ R_{n,\theta}(1) - \tilde{\mu}_\theta^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta}(1)} + \tilde{\mu}_\theta^+ R_{n,\theta,\epsilon}(1) \frac{\mu_\theta R_{n,\theta,\epsilon}(1) - \mu_\theta R_{n,\theta}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1) \mu_\theta R_{n,\theta}(1)} \right] \leq \frac{C\epsilon \tilde{\mu}_\theta^+(1)}{\mu_\theta R_{n,\theta}(1)} + C\epsilon \frac{\tilde{\mu}_\theta^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1) \mu_\theta R_{n,\theta}(1)}.$$

Now one can show that there exist a $C < +\infty$ such that for any $x, y \in \mathbf{X}$

$$R_{n,\theta}(1)(x) \geq CR_{n,\theta}(1)(y) \quad R_{n,\theta,\epsilon}(1)(x) \geq CR_{n,\theta,\epsilon}(1)(y). \quad (37)$$

Then it follows that

$$\frac{C\epsilon \tilde{\mu}_\theta^+(1)}{\mu_\theta R_{n,\theta}(1)} + C\epsilon \frac{\tilde{\mu}_\theta^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1) \mu_\theta R_{n,\theta}(1)} \leq C\epsilon \tilde{\mu}_\theta^+(1).$$

Hence we have shown that

$$\left[\frac{\tilde{\mu}_\theta^+ R_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)} - \frac{\tilde{\mu}_\theta^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)} \right] [F_\theta^{(n)}(\tilde{\mu}_\theta^+)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)] \leq C\|\varphi\|_\infty \epsilon \tilde{\mu}_\theta^+(1).$$

Second, the second line of the R.H.S. of (35). By Lemma D.2, for any $\mu_\theta \in \mathcal{P}(\mathbf{X})$, $\|F_\theta^{(n)}(\mu_\theta) - F_{\theta,\epsilon}^{(n)}(\mu_\theta)\| \leq C\epsilon$, with C independent of μ_θ , and in addition using (37) we have

$$\frac{\tilde{\mu}_\theta^+ R_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)} [F_\theta^{(n)}(\tilde{\mu}_\theta^+)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi) - F_{\theta,\epsilon}^{(n)}(\tilde{\mu}_\theta^+)(\varphi) + F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)] \leq C\|\varphi\|_\infty \epsilon \tilde{\mu}_\theta^+(1).$$

Thus we have shown:

$$\|\tilde{G}_\theta^{(n)}(\mu_\theta, \tilde{\mu}_\theta)(\varphi) - \tilde{G}_{\theta,\epsilon}^{(n)}(\mu_\theta, \tilde{\mu}_\theta)(\varphi)\| \leq C\epsilon[\tilde{\mu}_\theta^+(1) + \tilde{\mu}_\theta^-(1)] = C\epsilon\|\tilde{\mu}_\theta\|. \quad (38)$$

Control of $\tilde{H}_\theta^{(n)}(\mu_\theta)(\varphi) - \tilde{H}_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)$. We have

$$\tilde{H}_\theta^{(n)}(\mu_\theta)(\varphi) - \tilde{H}_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi) = \left[\frac{\mu_\theta \tilde{R}_{n,\theta}(\varphi)}{\mu_\theta R_{n,\theta}(1)} - \frac{\mu_\theta \tilde{R}_{n,\theta,\epsilon}(\varphi)}{\mu_\theta R_{n,\theta,\epsilon}(1)} \right] + \left[\frac{\mu_\theta \tilde{R}_{n,\theta,\epsilon}(1) F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)}{\mu_\theta R_{n,\theta,\epsilon}(1)} - \frac{\mu_\theta \tilde{R}_{n,\theta}(1) F_\theta^{(n)}(\mu_\theta)(\varphi)}{\mu_\theta R_{n,\theta}(1)} \right]. \quad (39)$$

We start with the first bracket on the R.H.S. of (39). We first note that

$$\tilde{R}_{n,\theta}(\varphi)(x) - \tilde{R}_{n,\theta,\epsilon}(\varphi)(x) = \int f_\theta(x'|x) \varphi(x') [\nabla g_\theta(y_n|x') - \nabla g_{\theta,\epsilon}(y_n|x')] dx' \leq C \|\varphi\|_\infty \epsilon \quad (40)$$

where we have applied (21). Then we have

$$\frac{\mu_\theta \tilde{R}_{n,\theta}(\varphi)}{\mu_\theta R_{n,\theta}(1)} - \frac{\mu_\theta \tilde{R}_{n,\theta,\epsilon}(\varphi)}{\mu_\theta R_{n,\theta,\epsilon}(1)} = \frac{\mu_\theta \tilde{R}_{n,\theta}(\varphi) - \mu_\theta \tilde{R}_{n,\theta,\epsilon}(\varphi)}{\mu_\theta R_{n,\theta}(1)} + \mu_\theta \tilde{R}_{n,\theta,\epsilon}(\varphi) \frac{\mu_\theta R_{n,\theta,\epsilon}(1) - \mu_\theta R_{n,\theta}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1) \mu_\theta R_{n,\theta}(1)}.$$

By using (40) on the first term on the R.H.S. of the above equation and by using (36) in the numerator for the second, along with (37) in the denominator, we have

$$\left| \frac{\mu_\theta \tilde{R}_{n,\theta}(\varphi)}{\mu_\theta R_{n,\theta}(1)} - \frac{\mu_\theta \tilde{R}_{n,\theta,\epsilon}(\varphi)}{\mu_\theta R_{n,\theta,\epsilon}(1)} \right| \leq C \epsilon [\|\varphi\|_\infty + |\mu_\theta \tilde{R}_{n,\theta,\epsilon}(\varphi)|].$$

Then as

$$\tilde{R}_{n,\theta,\epsilon}(\varphi)(x) = \int \varphi(x') [\nabla g_{\theta,\epsilon}(y_n|x') f_\theta(x'|x) - g_{\theta,\epsilon}(y_n|x) \nabla f_\theta(x'|x)] dx' \leq C \|\varphi\|_\infty \int_{\mathbf{X}} dx' \leq C \|\varphi\|_\infty \quad (41)$$

where the compactness of \mathbf{X} and (A5) have been used, we have the upper-bound

$$\left| \frac{\mu_\theta \tilde{R}_{n,\theta}(\varphi)}{\mu_\theta R_{n,\theta}(1)} - \frac{\mu_\theta \tilde{R}_{n,\theta,\epsilon}(\varphi)}{\mu_\theta R_{n,\theta,\epsilon}(1)} \right| \leq C \|\varphi\|_\infty \epsilon. \quad (42)$$

Moving onto the second bracket on the R.H.S. of (39), this is equal to

$$\left[\frac{\mu_\theta \tilde{R}_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)} - \frac{\mu_\theta \tilde{R}_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)} \right] F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi) + \frac{\mu_\theta \tilde{R}_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)} [F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)]$$

By using the inequality (42), we have

$$\left[\frac{\mu_\theta \tilde{R}_{n,\theta,\epsilon}(1)}{\mu_\theta R_{n,\theta,\epsilon}(1)} - \frac{\mu_\theta \tilde{R}_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)} \right] F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi) \leq C \epsilon |F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)| \leq C \|\varphi\|_\infty \epsilon.$$

Using Lemma D.2 and in addition using (37) in the denominator and (41) in the numerator we have

$$\frac{\mu_\theta \tilde{R}_{n,\theta}(1)}{\mu_\theta R_{n,\theta}(1)} [F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi) - F_\theta^{(n)}(\mu_\theta)(\varphi)] \leq C \|\varphi\|_\infty \epsilon$$

where C does not depend upon μ_θ and ϵ . Thus we have established that

$$\frac{\mu_\theta \tilde{R}_{n,\theta,\epsilon}(1) F_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)}{\mu_\theta R_{n,\theta,\epsilon}(1)} - \frac{\mu_\theta \tilde{R}_{n,\theta}(1) F_\theta^{(n)}(\mu_\theta)(\varphi)}{\mu_\theta R_{n,\theta}(1)} \leq C \|\varphi\|_\infty \epsilon. \quad (43)$$

One can put together the results of (42) and (43) and establish that

$$|\tilde{H}_\theta^{(n)}(\mu_\theta)(\varphi) - \tilde{H}_{\theta,\epsilon}^{(n)}(\mu_\theta)(\varphi)| \leq C \|\varphi\|_\infty \epsilon. \quad (44)$$

On combining the results (38) and (44) and noting (39) we conclude the proof. \square

Lemma D.2. Assume (A1-3). Then there exist a $C < +\infty$ such that for any $n \geq 1$, $\mu_\theta \in \mathcal{P}(\mathbf{X})$, $\epsilon > 0$, $\theta \in \Theta$:

$$\|F_\theta^{(n)}(\mu_\theta) - F_{\theta,\epsilon}^{(n)}(\mu_\theta)\| \leq C \epsilon.$$

Proof. For $\varphi \in \mathcal{B}_b(\varphi)$

$$F_{\theta}^{(n)}(\mu_{\theta})(\varphi) - F_{\theta,\epsilon}^{(n)}(\mu_{\theta})(\varphi) = \frac{\mu_{\theta}R_{n,\theta}(\varphi) - \mu_{\theta}R_{n,\theta,\epsilon}(\varphi)}{\mu_{\theta}R_{n,\theta}(1)} + \mu_{\theta}R_{n,\theta,\epsilon}(\varphi) \left[\frac{\mu_{\theta}R_{n,\theta,\epsilon}(1) - \mu_{\theta}R_{n,\theta}(1)}{\mu_{\theta}R_{n,\theta,\epsilon}(1)\mu_{\theta}R_{n,\theta}(1)} \right].$$

Then by applying (36) on both terms on the R.H.S. we have the upper-bound

$$\frac{C\|\varphi\|_{\infty}\epsilon}{\mu_{\theta}R_{n,\theta}(1)}.$$

One can conclude by using the inequality (37) for $R_{n,\theta}(1)(\cdot)$. \square

Lemma D.3. *Assume (A1-5). Then there exist a $C < +\infty$ such that for any $n \geq 1$, $\mu_{\theta} \in \mathcal{P}(\mathbf{X})$, $\widetilde{\mu}_{\theta} \in \mathcal{M}(\mathbf{X})$, $\epsilon > 0$, $\theta \in \Theta$:*

$$\|\widetilde{F}_{\theta}^n(\mu_{\theta}, \widetilde{\mu}_{\theta})\| \vee \|\widetilde{F}_{\theta,\epsilon}^n(\mu_{\theta}, \widetilde{\mu}_{\theta})\| \leq C(1 + \|\widetilde{\mu}_{\theta}\|).$$

Proof. We will consider only $F_{\theta}^n(\mu_{\theta}, \widetilde{\mu}_{\theta})$ as the ABC filter derivative will follow similar calculations, for any $\epsilon > 0$ (with upper-bounds that are independent of ϵ). By [29, Lemma 6.4] we have for $\varphi \in \mathcal{B}_b(\mathbf{X})$

$$\widetilde{F}_{\theta}^n(\mu_{\theta}, \widetilde{\mu}_{\theta})(\varphi) = \widetilde{G}_{\theta}^n(\mu_{\theta}, \widetilde{\mu}_{\theta})(\varphi) + \sum_{p=1}^n \widetilde{G}_{\theta}^{n-p}(F_{\theta}^p(\mu_{\theta}), \widetilde{H}_{\theta}^p(\mu_{\theta}))(\varphi).$$

By [29, Lemma 6.6] we have the upper-bound

$$\|\widetilde{F}_{\theta}^n(\mu_{\theta}, \widetilde{\mu}_{\theta})\| \leq C\left(\rho^n\|\widetilde{\mu}_{\theta}\| + \sum_{p=1}^n \rho^{n-p}\|\widetilde{H}_{\theta}^p(\mu_{\theta})\|\right)$$

with $\rho \in (0, 1)$. Then by [29, Lemma 6.8], it follows that

$$\|\widetilde{F}_{\theta}^n(\mu_{\theta}, \widetilde{\mu}_{\theta})\| \leq C\left(\rho^n\|\widetilde{\mu}_{\theta}\| + \sum_{p=1}^n \rho^{n-p}\right)$$

from which one concludes. \square

Remark D.1. *Using the proof above, one can also show that there exist a $C < +\infty$ such that for any $n \geq 1$, $\mu_{\theta} \in \mathcal{P}(\mathbf{X})$, $\widetilde{\mu}_{\theta} \in \mathcal{M}(\mathbf{X})$, $\epsilon > 0$, $\theta \in \Theta$*

$$\|\widetilde{F}_{\theta}^{(n)}(\mu_{\theta}, \widetilde{\mu}_{\theta})\| \vee \|\widetilde{F}_{\theta,\epsilon}^{(n)}(\mu_{\theta}, \widetilde{\mu}_{\theta})\| \leq C(1 + \|\widetilde{\mu}_{\theta}\|).$$

References

- [1] ANDRIEU, C., DOUCET, A. & TADIC, V. B. (2009). On-line simulation-based algorithms for parameter estimation in general state-space models, Technical Report, University of Bristol.
- [2] BARTHELMÉ, S. & CHOPIN, N. (2011). Expectation-Propagation for summary-less, likelihood-free inference. Technical Report, ENSAE.
- [3] BESKOS, A., CRISAN, D., JASRA, A. & WHITELEY, N. (2011). Error bounds and normalizing constants for sequential Monte carlo in high-dimensions. Technical Report, Imperial College London.
- [4] BICKEL, P., LI, B. & BENGTTSSON, T. (2008). Sharp failure rates for the bootstrap particle filter in high dimensions. In *Pushing the Limits of Contemporary Statistics*, B. Clarke & S. Ghosal, Eds, 318–329, IMS.
- [5] CAPPÉ, O., RYDEN, T. & MOULINES, É. (2005). *Inference in Hidden Markov Models*. Springer: New York.
- [6] CALVET, C. & CZELLAR, V. (2012). Accurate methods for approximate Bayesian computation filtering. Technical Report, HEC Paris.
- [7] CÉROU, F., DEL MORAL, P. & GUYADER, A. (2011). A non-asymptotic variance theorem for un-normalized Feynman-Kac particle models. *Ann. Inst. Henri Poincaré*, **47**, 629–649.

- [8] DEAN, T. A., SINGH, S. S., JASRA, A. & PETERS G. W. (2010). Parameter estimation for Hidden Markov models with intractable likelihoods. Technical Report, University of Cambridge.
- [9] DEAN, T.A. & SINGH, S.S. (2011) Asymptotic behaviour of approximate Bayesian estimators. Technical Report, University of Cambridge.
- [10] DEL MORAL, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer: New York.
- [11] DEL MORAL, P., DOUCET, A. & JASRA, A. (2006). Sequential Monte Carlo samplers. *J. R. Statist. Soc. B*, **68**, 411–436.
- [12] DEL MORAL, P., DOUCET, A., & JASRA, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statist. Comp.*, **22**, 1009–1020.
- [13] DEL MORAL, P., DOUCET, A. & SINGH, S. S. (2009). Forward only smoothing using Sequential Monte Carlo. Technical Report, University of Cambridge.
- [14] DEL MORAL, P., DOUCET, A. & SINGH, S. S. (2011). Uniform stability of a particle approximation of the optimal filter derivative. Technical Report, University of Cambridge.
- [15] DOUCET, A., GODSILL, S. & ANDRIEU, C (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statist. Comp.*, **10**, 197–208.
- [16] FREI, M. & KÜNSCH, H. (2012). Sequential state and parameter estimation using combined ensemble Kalman and particle filter updates. *Month. Weath. Rev.*, (to appear).
- [17] GAUCHI, J. P. & VILA, J. P. (2012). Nonparametric filtering approaches for identification and inference in nonlinear dynamic systems. *Statist. Comp.* (to appear).
- [18] JASRA, A., SINGH, S. S., MARTIN, J. S. & MCCOY, E. (2012). Filtering via approximate Bayesian computation. *Statist. Comp.*, (to appear).
- [19] LE GLAND, F. & MEVEL, M. (1997). Recursive identification in hidden Markov models. *Proc. 36th IEEE Conf. Decision and Control*, 3468–3473.
- [20] MARIN, J.-M., PUDLO, P., ROBERT, C.P., & RYDER, R (2012). Approximate Bayesian Computational methods. *Statist. Comp.*, (to appear).
- [21] MARTIN, J. S., JASRA, A., SINGH, S. S., WHITELEY, N. & MCCOY, E. (2012). Approximate Bayesian computation for smoothing, Technical Report, Imperial College London.
- [22] MCKINLEY, J., COOK, A. & DEARDON, R. (2009). Inference for epidemic models without likelihoods. *Intl. J. Biostat.*, **5**, a24.
- [23] MURRAY, L. M., JONES, E. & PARSLow, J. (2011). On collapsed state-space models and the particle marginal Metropolis-Hastings sampler. Technical Report, CSIRO.
- [24] NOTT, D., MARSHALL, L. & NGOC, T. M. (2012). The ensemble Kalman filter is an ABC algorithm. *Statist. Comp.*, (to appear).
- [25] PITT, M. K. (2002). Smooth particle filters for likelihood evaluation and maximization, Technical Report, University of Warwick.
- [26] POYIADJIS, G., DOUCET, A. & SINGH, S.S. (2011) Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, **98**, 65–80.
- [27] POYIADJIS, G., SINGH, S. S. & DOUCET, A. (2006). Gradient-free maximum likelihood parameter estimation with particle filters. *Amer. Control Conf.*, 6–9.
- [28] SPALL, J. (2003). *Introduction to Stochastic Search and Optimization* (1st ed), Wiley: New York.
- [29] TADIC, V. B. & DOUCET, A. (2005). Exponential forgetting and geometric ergodicity for optimal filtering in general state-space models. *Stoch. Proc. Appl.*, **115**, 1408–1436.

- [30] WHITELEY, N., KANTAS, N. & JASRA, A. (2012). Linear variance bounds for particle approximations of time-homogeneous Feynman-Kac formulae. *Stoch. Proc. Appl.*, **122**, 1840–1865.
- [31] YILDIRIM, S., SINGH, S.S. & JASRA, A. (2012). Expectation Maximisation for approximate Bayesian computation maximum Likelihood estimation. Technical Report, University of Cambridge.